communications biology

REVIEW ARTICLE

https://doi.org/10.1038/s42003-021-02586-0 OPEN

Accelerating antibiotic discovery through artificial intelligence

Marcelo C. R. Melo [1,2,3,5], Jacqueline R. M. A. Maasch [1,2,3,4,5] & Cesar de la Fuente-Nunez [1,2,3]✉

By targeting invasive organisms, antibiotics insert themselves into the ancient struggle of the host-pathogen evolutionary arms race. As pathogens evolve tactics for evading antibiotics, therapies decline in efficacy and must be replaced, distinguishing antibiotics from most other forms of drug development. Together with a slow and expensive antibiotic development pipeline, the proliferation of drug-resistant pathogens drives urgent interest in computational methods that promise to expedite candidate discovery. Strides in artificial intelligence (AI) have encouraged its application to multiple dimensions of computer-aided drug design, with increasing application to antibiotic discovery. This review describes AI-facilitated advances in the discovery of both small molecule antibiotics and antimicrobial peptides. Beyond the essential prediction of antimicrobial activity, emphasis is also given to antimicrobial compound representation, determination of drug-likeness traits, antimicrobial resistance, and *de novo* molecular design. Given the urgency of the antimicrobial resistance crisis, we analyze uptake of open science best practices in AI-driven antibiotic discovery and argue for openness and reproducibility as a means of accelerating preclinical research. Finally, trends in the literature and areas for future inquiry are discussed, as artificially intelligent enhancements to drug discovery at large offer many opportunities for future applications in antibiotic development.

Antimicrobial resistance (AMR) in clinically significant bacteria is undermining the efficacy of existing antibiotics, incurring concerning levels of global morbidity and mortality[1]. The Centers for Disease Control and Prevention estimates that 2.8 million infections are caused by antibiotic-resistant bacteria in the United States annually, leading to 35,000 deaths from such untreatable infections[2]. Current evidence also suggests that the solution may be part of the problem itself: antibiotics have been shown to cause significant damage to the gut microbiome, reducing species diversity and encouraging the evolution and dissemination of AMR genes[3]. Antibiotics under clinical trial are generally analogs to existing drugs for which AMR mechanisms have already emerged[1], further underscoring the need for novel approaches in antibiotic discovery.

Compounding this issue, antibiotic development is a slow, expensive, and failure-prone process that can span over 10 years and cost hundreds of millions of dollars[4]. Between 2014 and 2019, only 14 new antibiotics were developed and approved[5]. In a survey of nearly 186,000 clinical trials for over 21,000 compounds, the probability of success for new drugs that treat

[1] Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2] Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [3] Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA, USA. [4] Department of Computer and Information Science, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [5] These authors contributed equally: Marcelo C. R. Melo, Jacqueline R. M. A. Maasch. ✉email: cfuente@upenn.edu

infectious diseases was 25.2%[6]. For orphan drugs, i.e., those that treat rare infectious diseases, this probability dropped to only 19.1%[6]. This risk of failure drives corporations to pursue research and development with a higher guarantee of return on investment, opening the way for academia to initiate early stages of antibiotic design and optimization[7,8].

Accelerated antibiotic discovery will require computer-aided prospection for novel drugs with new mechanisms of action (MOAs)[9]. It is speculated that 10^{30}–10^{60} drug-like chemicals exist[10], while 20^n variants exist per n-length canonical amino acid sequence. Although this immense combinatorial space presents a broad opportunity for computational antibiotic design, an exhaustive search cannot be achieved on a reasonable timescale. These challenges strongly incentivize the development of efficient heuristics and artificially intelligent algorithms for high-throughput antibiotic discovery. A prominent subdomain of computer science, artificial intelligence (AI) concerns the study and development of machines that are capable of learning, problem-solving, or mimicking other displays of reasoning akin to natural intelligence. For the purposes of this review, AI will generally pertain to machine learning (ML), the training of mathematical models to output predictions when presented with previously unseen data. The application of ML to drug discovery, and antibiotic discovery specifically, has been greatly facilitated by the public availability of empirical datasets (Table 1), advances in computer engineering, and the proliferation of free and open-source ML libraries.

The integration of computational tools to expedite drug development has led to key advances for the rational design of bioactive compounds efficacious in animal models, thus demonstrating that computers can yield preclinical antibiotic candidates[9,11]. Leveraging advances in protein structure prediction and modeling, small-molecule antibiotic targets can be reliably described in atomic detail. Protein structures are then probed for binding sites, allowing large libraries of compounds to be used for automated large-scale docking and binding affinity studies in a process known as virtual screening (VS). This practice is now integral to many drug development pipelines, receiving ample attention from the ML community[12]. The most challenging step in VS is evaluating binding site affinity, driving the development of ML tools that significantly outperform traditional binding affinity prediction methods[13–15]. In recent years, deep learning (DL) has been used to successfully bypass docking and affinity estimation entirely, resulting in the identification of a small-molecule antibiotic active against multiple bacterial pathogens[16].

In this review, we will focus on the application of AI to the development of two major classes of bioactive compounds: small-molecule antibiotics and antimicrobial peptides (AMPs). The former, studied since the beginning of the twentieth century with the discovery of penicillin and in use for over 70 years, represents the majority of antibiotics in use today. The latter, a class of small proteins usually composed of 5 to 50 amino acids, is receiving increasing attention in research and clinical trials[17] due in part to a relatively low propensity to induce AMR[18]. Research topics will be introduced by following the logical flow of an ML pipeline, starting with compound representation and progressing through trait prediction and novel compound design. ML innovation in general drug development will be reviewed where it has cross-over utility for antibiotic-specific applications. Trends in the literature and directions for future research will be discussed, including prospects for increasing data availability, computational–experimental collaboration, and innovation in interpretable ML (IML). Additionally, we provide an original analysis of open science practices among cited

Table 1 Databases for computational antibiotic discovery.

Database	Site
General drug discovery and biomolecular informatics	
Binding MOAD[160]	https://bindingmoad.org
BindingDB[161]	https://www.bindingdb.org/
BRENDA[162]	https://www.brenda-enzymes.org
ChEMBL[163]	https://www.ebi.ac.uk/chembl/
Drug Design Data Resource	https://drugdesigndata.org
Drug Repurposing Hub[140]	https://clue.io/repurposing
DrugBank[164]	https://go.drugbank.com
MoleculeNet[165]	http://moleculenet.ai
Protein Data Bank[166]	https://www.wwpdb.org
PubChem[167]	https://pubchem.ncbi.nlm.nih.gov
Search Tool for Interacting Chemicals[168]	http://stitch.embl.de
Side Effect Resource[169]	http://sideeffects.embl.de
SuperTarget[170]	http://insilico.charite.de/supertarget/
Therapeutics Data Commons	https://zitniklab.hms.harvard.edu/TDC
Therapeutic Target DB[171]	http://db.idrblab.net/ttd/
UniProt[172]	https://www.uniprot.org
ZINC[173]	https://zinc15.docking.org
Exclusively infectious disease	
ADAM[174]	http://bioinformatics.cs.ntou.edu.tw/adam/
ADAPTABLE[175]	http://gec.u-picardie.fr/adaptable
Collection of Antimicrobial Peptides[176]	http://www.camp.bicnirrh.res.in
Data Repository of Antimicrobial Peptides[177]	http://dramp.cpu-bioinfor.org
DB of Antimicrobial Activity and Structure of Peptides[178]	https://dbaasp.org
dbAMP[179]	http://140.138.77.240/~dbamp
MEGARes: Antimicrobial DB for High-Throughput Sequencing[180]	https://megares.meglab.org
National DB of Antibiotic-Resistant Organisms	https://www.ncbi.nlm.nih.gov/
Pathosystems Resource Integration Center[181]	https://www.patricbrc.org
Tropical Disease Research Targets[182]	https://tdrtargets.org

Public databases (DB) of general use in computational drug discovery and biomolecular informatics, as well as those specific to antimicrobial discovery and resistance.

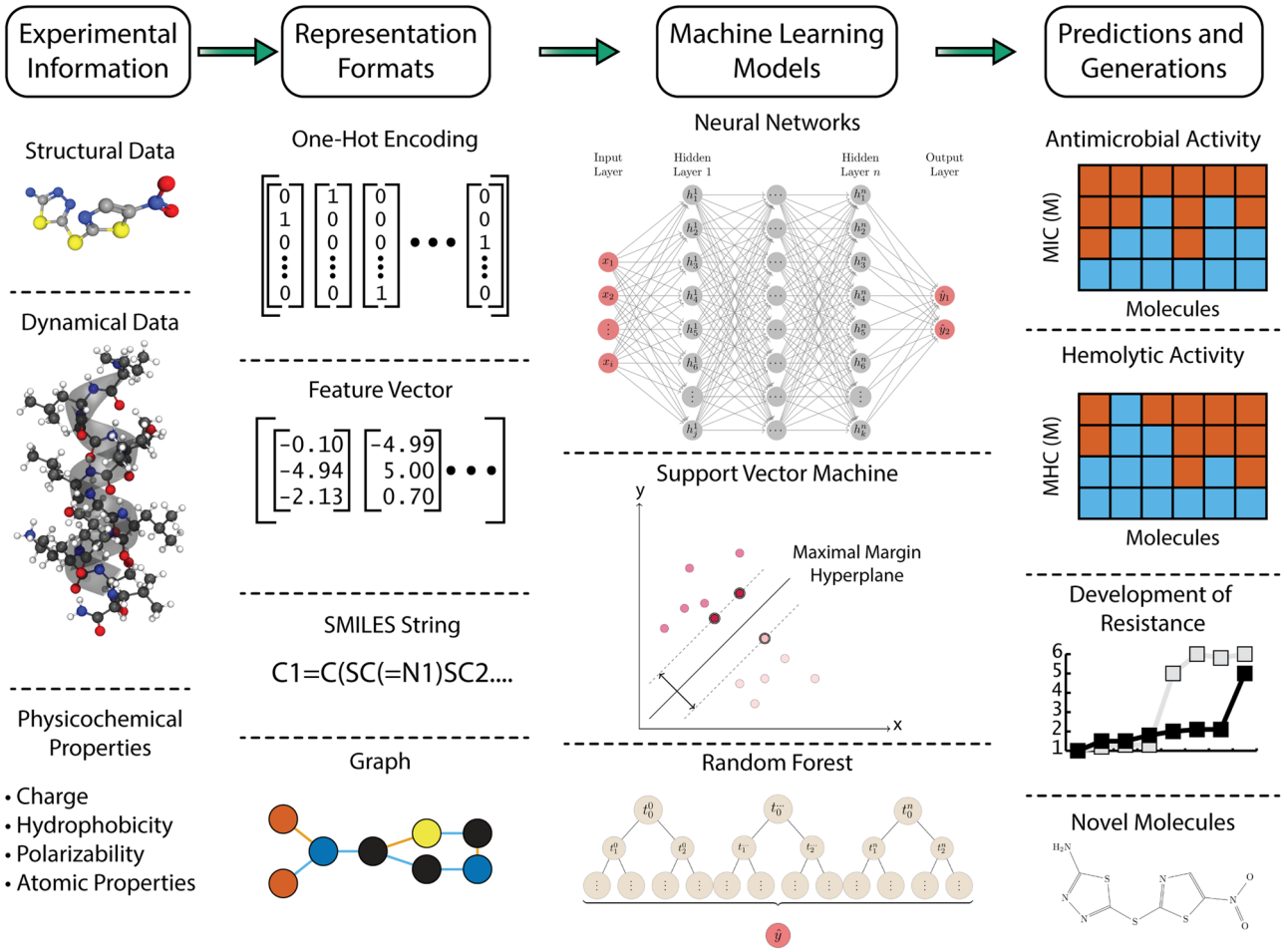

Fig. 1 Computational antibiotic discovery pipeline. The figure provides an overview of data and methods used in antibiotic discovery and development using AI. From left to right, key elements in the drug development process are exemplified. The first part of any AI-driven project is gathering the experimental information that will enable model creation. The data are then transformed into AI-ready representations. Subsequently, models are trained using algorithms that can range from traditional decision trees to novel neural networks. Finally, trained models can be used to predict diverse qualities, e.g., the effectiveness of an antibiotic, potential for toxic activity, development of resistance, or the structure of novel compounds that exhibit desirable traits.

research and discuss the potential for best practices in open and reproducible ML to expedite antibiotic discovery.

Methods for optimizing compound representation

The search for optimal measurements of quantitative structure–activity relationships (QSAR) drove over 50 years of research and innovation[19]. Aiming to computationally predict the activity of newly designed molecules, saving time and money by avoiding synthesis and experimentation on inactive compounds, researchers relied on computational representations of drug candidates to predict their properties. As it became apparent, the problem of representing biological or chemical data for use in computational models is in itself an important field of research. Likewise, it is an essential component of the computational drug discovery pipeline (Fig. 1). The variety of information sources and experimental procedures to describe molecules can rapidly lead to overwhelming amounts of information, which may cause more harm than good. For example, in order to describe simple amino acid residues, over 400 different measurements have been performed and combined in online databases[20]. For small-molecule drugs, approaches range from calculating and condensing quantum mechanically derived descriptors[21] to calculating topological properties[22,23]. The sheer amount of data and the redundant information contained in multiple measurements makes using all descriptors impractical or counterproductive. This led to a series of studies that combined experimental data into reduced descriptors that maximized information content in as few dimensions as possible[24].

From traditional dimensionality reduction techniques like principal component analysis (PCA) and singular value decomposition to feature selection approaches involving χ^2 statistical tests or mutual information estimation, the search for reduced and information-rich representations has now fully integrated ML tools and principles. The efforts described below highlight how the theoretical and methodological advances made in diverse ML applications and interest areas can be adapted to aid ML-driven antibiotic development.

A prominent example is the use of graph convolutional networks to leverage the geometry and connectivity of molecules to naturally translate them into graphs, using neural networks to learn from the chemical structure itself[25]. A similar approach was taken to study and predict protein structures[26]. In an extensive benchmark study of available methods and datasets[27], it was found that neural networks can enhance not only the process of describing a drug based on a set of molecular descriptors but also

the determination of such molecular descriptors themselves. This work was extended to create a series of antimicrobial compounds that were correctly predicted as active despite being structurally distant from known antibiotics [16].

While common in the fields of signal processing and natural language processing (NLP), recurrent neural networks (RNNs) have now been adapted to process simplified molecular-input line-entry system (SMILES) representations, which encode structures of chemical species using simple text strings. In one case, researchers used long short-term memory (LSTM) generative neural networks to learn from SMILES representations of known drugs and then used the trained neural network to generate new compounds[28]. Alternatively, RNNs have been combined with reinforcement learning to autonomously create an embedded representation for drugs based on their SMILES representations[29].

RNNs have emerged as a natural embedding approach for AMP sequences, given their ability to parse sequence-based inputs. Based on a one-hot encoding of amino acid residues (i.e., a 20-mer vector with 19 "zeros" and a single "one" at unique positions to indicate different residues), both an LSTM-based autoencoder[30] and multiplicative-LSTM neural network[31] have been trained to create embedded representations for peptide sequences. The latter led to an embedded representation that could be used to derive a protein's secondary structure, thermal stability, deep mutational scanning classification, and even the functional impact of mutations[31].

Antimicrobial activity prediction

Predicting antimicrobial activity is at the core of ML integration into antibiotic development, driving over 10 years of research to provide new solutions for the QSAR problem[7] and attracting a variety of approaches[32] (Table 2). For instance, to improve upon previous attempts to design new drugs based on the analysis of chemical fragments and their properties, researchers used multinomial logistic regression to classify fragments that comprise molecules in a training set. This process created a "vocabulary" of fragments that could then be combined to propose new antibiotics active against the Gram-negative bacterium *Pseudomonas aeruginosa*[33].

In a recent effort to repurpose previously developed drugs as antibiotics[16], a combination of neural network models was used to create a new representation for chemical compounds, and then assess their antimicrobial potential. Interestingly, this effort also made use of *ensemble learning*[34], a technique that combines multiple copies of a model (with different weights or architectures) and takes a weighted vote of each model into consideration to achieve the final prediction[35]. The underlying assumption behind ensemble learning is that errors made by one model will be compensated for by others, and this assumption has been confirmed in applications ranging from proinflammatory peptide identification[36] to prediction of drug side effects[37].

Classical ML techniques such as support vector machines (SVMs) have been applied to describe AMPs and quantify their MOAs[38,39]. Alternatively, deep neural networks have been used to predict antimicrobial properties from simplified residue representations of arbitrary amino acid sequences. In 2009, researchers combined 44 peptide descriptors traditionally used for QSAR studies and used them as inputs for an artificial neural network that predicted peptide activity against *P. aeruginosa*[40]. More recently, a 2020 study created a deep convolutional neural network model based on a simplified amino acid vocabulary that translated the natural 20 amino acids into pseudo residue types[41]. This model predicts antimicrobial activity in small peptides and is available in a web server. Extreme gradient boosting has been used for genome-based prediction of minimum inhibitory concentrations for 20 antibiotics against *Klebsiella pneumoniae*[42] and 15 antibiotics against nontyphoidal *Salmonella* strains[43]. Using RNNs[44], a combination of input representation and regression

Table 2 Machine learning models for antibiotic discovery.

	Public release			
Algorithm	**Code**	**Data**	**Software**	**Software type**
Antimicrobial activity prediction				
Artificial neural network[40]		Yes		
Support vector machine[38]		Yes		
Multinomial logistic regression[33]		Yes		
LSTM RNN[44]	Yes	Yes	Yes	Command-line tool
XGBoost[42]	Yes	Yes	Yes	Command-line tool
Directed-message passing neural network[16]	Yes	Yes	Yes	Web server, Docker container
DBSCAN[47]		Yes	Yes	Web server
DBSCAN[48]			Yes	Web server
Convolutional neural network[41]		Yes	Yes	Web server
Generalized linear model[49]				
Random forest[50]				
Hemolytic activity prediction				
Classification and regression trees[55]		Yes		
Artificial neural network[54]		Yes	Yes	Web server
Gradient boosting classifiers[56]	Yes	Yes		
Support vector machine[183]		Yes	Yes	Web server, mobile app, standalone
De novo antibiotic design				
Variational autoencoder[45]		Yes		
LSTM RNN[30]	Yes	Yes	Yes	Command-line tool
LSTM RNN[120]				
Generative adversarial network[119]	Yes	Yes	Yes	Command-line tool

Machine learning models cited in this review pertain specifically to antimicrobial compound discovery, i.e., those that predict antimicrobial activity, those trained on antimicrobial compound data to predict drug-likeness, and those that generate potential antimicrobials. Public release of model source code, training and/or testing data, and/or associated software tools are noted. Criteria for data release were lenient, with "yes" indicating partial or full release of training or testing data.

models were created to select peptide sequences with antimicrobial activity. Finally, through a variational autoencoder approach, peptide sequences were embedded in a latent space that was subsequently searched for new AMP sequences[45].

The variety of techniques utilized thus far correlates with an increasing focus on AMPs, which have been regarded as a major source of new antibiotics to tackle the development of resistance in microbes[9]. The ability of AMPs to limit AMR development has been related to their varied MOAs[46], which has led researchers to focus on classifying peptides and discovering new MoAs. Specifically, DBSCAN was used for cluster-based prediction of AMP activity against Gram-negative bacteria[47], with promising candidates being synthesized and tested in vitro[48].

The direct combination of experimental and ML techniques in a closed-loop approach has also benefited the development of new AMPs. Starting from a template with known antimicrobial activity and a series of homologous sequences, it was possible to train a generalized linear model to create new AMPs with 160-fold increased antimicrobial activity against *Escherichia coli*[49]. Since patterns found by generalized linear models can be directly interpreted by analyzing the model weights, one can directly translate the model into actionable information for AMP design.

While most ML-based antibiotic development approaches focus on creating new representations for drug candidates and new models to predict their activity based on molecular descriptors, the phenotypic drug discovery approach focuses not on describing the molecule itself but on its effects on target organisms. For example, a recent study used a random forest model to predict antimicrobial activity based on featurization of cell imaging, avoiding detailed description of the molecules themselves[50]. This approach can expand the search space for new drugs by avoiding direct comparisons between molecular descriptors and focusing instead on their effects on pathogens.

Drug-likeness prediction

ML can yield a fuller aggregate picture of antibiotic therapeutic potential than simply predicting antimicrobial activity. Attempts to quantitatively distinguish the subsets of chemical space that have therapeutic potential from those which do not have yielded various schemas, including the introduction of the Rule of 5 in 1997[51] and subsequent concepts of drug-likeness and lead-likeness. Prediction of drug-likeness has been refined and increasingly automated over recent decades, with traits of interest including absorption, distribution, metabolism, excretion, and toxicity (ADMET)[10,52]. ML-based prediction of binding affinity can also accelerate high-throughput screening and structure-based drug lead optimization by pinpointing candidates with more favorable drug–target interactions, as discussed in recent reviews[15,53].

Like many ML problems, drug-likeness prediction can be attempted using a wide array of algorithms. While experimental observations often require a specific methodology or well-established gold standard, diverse ML algorithms can often provide comparable performance for a given classification or regression problem. There is often no way to know a priori which algorithm will perform best, although theoretical knowledge can guide decision-making. Therefore, it is important to follow a rigorous model selection process that compares multiple algorithms (e.g., a Gaussian process, random forest, SVM, and neural network) across several performance metrics that are salient to the particular use case. In this section, we note the use of diverse algorithms that have been applied to multiple drug-likeness prediction problems.

Dangerous pharmacokinetic properties and toxicity are leading causes of clinical trial failure[52], incentivizing pre-trial in silico exploration. Host cell toxicity is a critical ADMET endpoint and a significant risk in antibiotic development, motivating the design of predictive tools for mammalian red blood cell toxicity, kidney cell toxicity, and other forms of eukaryotic cell damage. Hemolytic activity, or the ability to burst red blood cells, has been a major focus of therapeutic development given that numerous drugs enter the bloodstream. Prediction of hemolytic activity in AMPs and antimicrobial peptidomimetics has been explored using neural networks[54], classification trees[55], and gradient boosting classifiers[56]. Consensus model-based software for hemolytic activity prediction has also been released for general applications in drug development, with an emphasis on small molecules[57] and saponins[58]. A feedforward fully connected neural network has demonstrated comparable performance to prior random forest models for the prediction of drug candidate cytotoxicity[59]. Deep Taylor Decomposition was used to identify the most significant features in DL-based cytotoxicity classification, with an emphasis on visualization to facilitate interpretability[59]. Additional antibiotic side effects can also be foreshadowed using ML, as has been done for the seizure-inducing potential of enoxacin, a broad-spectrum fluoroquinolone antibacterial[60].

The development of AMP-based antibiotics must also consider peptide solubility and stability, which are necessary for manufacture and efficacy. Pharmaceutically viable AMPs will be soluble, a trait that can be predicted from amino acid sequence[61]. Protein solubility prediction has used neural network[61,62], gradient boosting machine[63], logistic regression classifier[64], SVM[65], and random forest models[66]. Degradation via the action of proteolytic enzymes is a significant concern when evaluating the stability of peptide-based antibiotics[67,68]. The in silico identification of putative proteolytic cleavage sites can inform AMP lead selection and guide sequence optimization for increased stability. Cleavage site prediction has been explored through the lens of drug development[69] and other protein informatics applications using classification and regression mode SVM[70–74], convolutional neural network[75], conditional random field classifier[76], and logistic regression models[77]. Similarly, the stability of drug-like chemicals has been modeled using an attention-based graph convolution neural network[78] and Naive Bayes classifier[79].

As outliers to original drug-likeness definitions expand the boundaries of these criteria, new qualitative endpoints and quantitative thresholds have come under consideration[80]. Collateral damage to the gut microbiome has been proposed as one additional ADMET endpoint, and consensus model-based software has been released for ML prediction of microbiome damage[58]. Indeed, disruption to the microbiome is a significant side effect of antibiotics and has been implicated in AMR evolution[3]. For this particular endpoint, species-specific antimicrobial activity prediction may be the answer: ML can aid in the selection of candidates with high specificity for target pathogens and low activity against known commensals.

AMR prediction

Unlike most therapeutics, antibiotics are designed to kill a living target with the capacity for resistance evolution. The near-inevitability of AMR evolution thus adds an additional urgent consideration that is absent from most other drug development niches. Incentives to develop less resistance-prone countermeasures are drawing research to historically underexplored sources of inspiration for novel antibiotic design[9]. Likewise, the need to track AMR emergence, mechanisms, and dynamics are raising new applied ML questions unique to computational antibiotic discovery, bacterial genomics, and infectious disease epidemiology. While ML-based AMR prediction may be clinically useful for informing AMR diagnosis and antibiotic

prescription[81,82], it may also be experimentally useful in the drug development process. We anticipate that ML approaches to AMR genomics in epidemiology and medicine will increasingly be adapted specifically for drug development purposes, e.g. ML-informed resistance evolution experiments for new lead compounds.

Protein space is one such underexplored area that is expected to yield future antibiotics with minimal AMR risk. Antimicrobial host defense peptides, including encrypted AMPs released from precursor proteins through proteolytic cleavage, have notably emerged as reservoirs for low AMR-risk antibiotic templates due in part to a tendency to act on multiple cellular targets[18,46,83,84]. Small-molecule AMR has also been observed to coincide frequently with a collateral sensitivity to AMPs, yet rarely with AMP cross-resistance[85]. Together with the fact that protein target modifications are a common AMR mechanism, this suggests the large potential for cross-over between ML and traditional protein informatics in AMR research. However, the majority of existing ML models forgo this route in favor of pathogen genetic and genomic inputs. Although model design strategies are expected to diversify, the current state of ML for AMR learns from the bacterial genome rather than drug or molecular target features.

Pathogen genomic data have been used to build ML models of antibiotic susceptibility and resistance phenotypes in clinically relevant bacteria, including *K. pneumoniae*[42], *E. coli*[86–88], *P. aeruginosa*[86,89], *Mycobacterium tuberculosis*[90,91], and *Staphylococcus aureus*[86]. While ML models of AMR may be trained on drug- and bacteria-specific data[92–94], a more agnostic approach has been explored using a neural network to facilitate environmental metagenomic analysis[95]. However, predictive performance has been observed to vary significantly by antibiotic, target species, genomic data sampling method, and resistance mechanism complexity[82,96], suggesting that AMR prediction may at times require relatively context-specific modeling. A free web server and standalone software have been released for SVM-based prediction of efflux-mediated AMR[97]. ML-assisted metagenomic analysis has implicated AMR genes associated with antibiotic-induced microbiome perturbations[98]. A novel combination of protein homology modeling and LASSO penalized logistic regression has been used to investigate the horizontal transfer of antibiotic resistance determinants from gut commensals to bacterial pathogens[99].

While "black-box" approaches may limit the utility of ML for AMR-risk reduction[81], IML can enable models to suggest causal factors in AMR at the organismal and population scale. Coupling ML with gene–protein structure mapping to investigate drivers of *M. tuberculosis* AMR evolution, interactions between genes conferring AMR were hypothesized to manifest as correlations in their weights and signs across the hyperplanes of an SVM ensemble[100]. An ML-integrated genome-scale model using data from microbial genome-wide association studies has enabled allele-parameterized flux balance analysis to reveal metabolomic insights into *M. tuberculosis* AMR[101]. Open-source software using protein orthology-based gene variant mapping has also been developed for interpretable AMR prediction[96]. Computationally characterizing the molecular signatures and population dynamics of AMR might help indicate which MOAs are overused and which present promising new avenues, even on a regional scale. Using training data from multiple countries, geographic analysis of predicted AMR genes revealed population dynamics that could be supported by national rates of multidrug-resistant tuberculosis and antibiotic prescription trends[100].

Generative DL for antibiotic discovery

Generative DL can lend itself to computational antibiotic discovery in multiple ways. Here, we will focus on de novo molecular design, which often employs generative adversarial networks (GANs), variational autoencoders (VAEs), or related architectures. Comprised of dueling generative and discriminative models, GANs infer the probability distribution from which training data derive in order to construct novel samples from this distribution. Engaging in a two-player minimax game, both models are trained to optimize the error rate of the discriminator: while the generator is trained to maximize the likelihood that the discriminator fails to distinguish empirical data from synthetic data, the discriminator is trained to minimize this likelihood[102]. Like classical autoencoders, VAEs are trained to encode inputs to a compressed representation and then to decode an approximate reconstruction, learning the latent variables describing the training data in the process. However, VAEs are directed probabilistic models, learning continuous latent variables through a variational Bayesian approach to generative DL[103]. This section will note the use of several variations on these common generative architectures as applied to drug discovery.

Generative DL has found diverse chemical and protein engineering applications[104], including inverse design of inorganic matter[105] and graph-based neural network models for the NP-hard[106] inverse protein folding problem[26,107]. Increasingly, generative DL is applied explicitly to drug discovery, whereby synthetic molecular designs are proposed from drug-like chemical spaces. De novo drug candidate design has been attempted with deep reinforcement learning coupling generative and predictive neural networks[29], deep generative adversarial autoencoder architecture[108], differentiable neural computer architecture with reinforcement learning and adversarial training[109,110], deep neural networks coupled with Monte Carlo tree search[111], and an autoencoder–GAN combination for both random and target-biased molecular design[112]. Given their suitability for sequential data, generative RNNs taking SMILES inputs have drawn attention in drug design[113,114] and have demonstrated relatively broad, uniform, and complete coverage of chemical space[115,116]. Experimentally validated membranolytic anticancer peptides have been generated by both an LSTM RNN with transfer learning[117] and a counterpropagation artificial neural network optimized by a genetic algorithm[118].

A burgeoning interest in generative DL within chemical engineering, protein engineering, and drug development at large suggests that similar techniques may be increasingly applied to AMP and small-molecule antibiotic design. To date, a GAN has been used to generate an AMP with a significantly lower minimum inhibitory concentration against *E. coli* than ampicillin[119]. Additional preliminary success in AMP discovery is described in a proof-of-concept study coupling a VAE with experimental validation[45]. A generative LSTM RNN with transfer learning has demonstrated success in reconstructing molecules known to target *S. aureus* after pretraining on a large generalized dataset and fine-tuning on a smaller set of target-specific bioactive molecules[120]. An RNN with unidirectional LSTM cells for de novo AMP design observed 82% of generated peptides to be putative AMPs, while only 65% of random permutations from the amino acid distribution of the training data were predicted to be antimicrobial[30].

Openness and reproducibility

In this section, we present an argument for increasing openness and reproducibility in ML-based antibiotic discovery. This argument hinges on a two-pronged crisis: (1) the global public health crisis of AMR, slow antibiotic development rates, and emerging infectious diseases and (2) the reproducibility crisis currently plaguing AI. We conclude with an original analysis of open science practices among the publications cited in this review.

Accelerating antibiotic discovery through open information and technology exchange carries both practical and ethical weight. As evidenced by poignant examples from the COVID-19 pandemic, factors such as AMR[121], sudden pathogen emergence, unexpected large-scale losses in quality of life and economic security[122], and structural inequities that render some populations disproportionately vulnerable[123] raise unique questions of urgency and justice in infectious disease control. These questions heighten the need for swift research and development, evoking calls for increased openness under global public health crises[124]. We argue that similar calls should extend to the global crisis of AMR evolution, and thus to computational antibiotic discovery.

The international movement toward open-access publishing represented by groups like cOAlition S[125] signal a growing concern for transparency, reproducibility, and equitable access to information within the scientific community. Effective 2021, Plan S dictates that publications resulting from public and private grants of participating bodies "must be published in Open-Access Journals, on Open-Access Platforms, or made immediately available through Open-Access Repositories without embargo" (https://www.coalition-s.org/plan_s_principles/). Nevertheless, open-access publishing addresses only one facet of computational openness and reproducibility. With stakes as high as they are in computational antibiotic discovery, we call for a more comprehensive set of open science best practices.

An open science regime that ensures computational reproducibility can accelerate ML-based antibiotic discovery through free public access to (1) source code, (2) training and testing data, and (3) published findings. Computational reproducibility facilitates the external validation of published claims while encouraging the dissemination of knowledge and methods. However, standards of openness and reproducibility in biomedical ML are still subject to debate[126], and some argue that AI generally suffers from a reproducibility crisis, not unlike that of psychology[127]. Reproducibility challenges common to ML (e.g., verbal descriptions in lieu of source code omitting essential hyperparameter values or random state seeds) can also have detrimental interactions with challenges unique to biomedicine (e.g., patient privacy laws precluding data sharing)[128].

Although releasing source code, training data, and testing data could mitigate reproducibility concerns while increasing the scientific value of AI research[126], an analysis of 400 general AI conference papers revealed that only 6% released code, 54% released pseudocode, and ~30% released test data[127]. Within ML for the life sciences and medicine specifically, a recent review found that 50% of 300 publications released software, while 64% released data[129]. A review of 511 studies found that papers applying ML to healthcare data underperformed relative to NLP, computer vision, and general ML on multiple metrics of reproducibility, including code release rates[130]. A systematic review of 415 studies on ML-based image analysis for COVID-19 diagnosis found that all publications contained serious methodological flaws or failed to report key information needed for reproducibility and substantiation of claims, such that not a single model was of clinical use[131].

Confounding factors such as lack of incentives in academia or misaligned objectives in the private sector may further hinder the adoption of open science practices. While open-access journals continue to grow, many prestigious scientific journals charge premiums over publication fees in order to make articles open access. Authors then face a tough choice between funding their research or paying premiums to make their publications free to all readers. Indeed, a recent study showed that authors of open-access publications in US research institutions tend to have more access to funding and belong to more advanced career stages[132]. Exemplifying the conflict between researchers and publishers, the recent 2-year-long negotiation between the University of California system and Elsevier resulted in the largest deal for open-access publishing for scientific articles in North America[133]. Interestingly, the fields of medical and biological research are among the most accessible, with biology having the largest fraction of immediately free-to-read articles[134].

Beyond publishing research findings, the release of source code and training and testing data may also raise conflicts regarding intellectual property (IP) and competitiveness in the private sector. Therefore, while industry-funded research for antimicrobial discovery[135] can still provide great advances to the field, finding a balance between open access and closed IP may prove to be a barrier in itself. Guidance may be found in the efforts of related fields to establish community-wide standards for responsible and reproducible ML publications, with the Checklist for Artificial Intelligence in Medical Imaging being a notable example[136].

This conversation in AI, and in biomedical ML specifically, motivated our analysis of code, data, and software release rates among models cited in this review (Fig. 2). This analysis was performed post hoc, such that all studies previously cited in this review that presented an ML model designed for antimicrobial compound discovery were included. It, therefore, focuses on key contributions in ML-facilitated antibiotic discovery, rather than an exhaustive literature analysis. Among ML models pertaining specifically to antimicrobial compound discovery (Table 2), we found that 31.6% (6/19) released code, 52.6% (10/19) released software, and 78.9% (15/19) released some or all training or testing data. Further, 26.3% (5/19) released code, data, and software, while 15.8% (3/19) released nothing in these three categories. It should be noted that our criteria for data release were lenient, with "yes" indicating partial or full release of training or testing data. Although best practice is to release full, metadata-documented versions of both training and testing datasets in a manner that is easily accessible for the reader, this is often not the standard followed in past publications. While our sample size is small, we hope that these statistics will inspire increased best-practice public release rates in ML for antibiotic discovery.

Moving forward, inspiration can be found in projects taking a broad view of openness and reproducibility in drug discovery. The open-source Therapeutics Data Commons (https://zitniklab.hms.harvard.edu/TDC/) provides free ML datasets to lower barriers to entry and accelerate drug development pipelines. The Open-Access Antimicrobial Screening Program extends the concept of openness to experimental methods by offering free compound screening services (https://www.co-add.org). Such creative counterexamples to the closed research paradigm will ideally become the norm in antibiotic discovery.

Trends and future directions

In this section, we examine research trends and discuss future trajectories for ML-facilitated antibiotic discovery. We anticipate that a trickle-down effect from adjacent ML research will stimulate significant AI-facilitated innovation in antibiotic discovery over the next decade. We expect this innovation process to require increased data quality and availability, exploration of new regions in chemical space, re-exploration of known regions through drug repurposing, collaboration between computational scientists and experimentalists, and enhanced explainability through IML.

To assess the state of publishing on ML for antibiotic discovery, we measured trends among papers in PubMed, a public database maintained by the United States National Library of Medicine of the National Institutes of Health (https://pubmed.ncbi.nlm.nih.gov). To explore the extent to which research interest has changed over time, we queried PubMed by year for texts on ML and antibiotics,

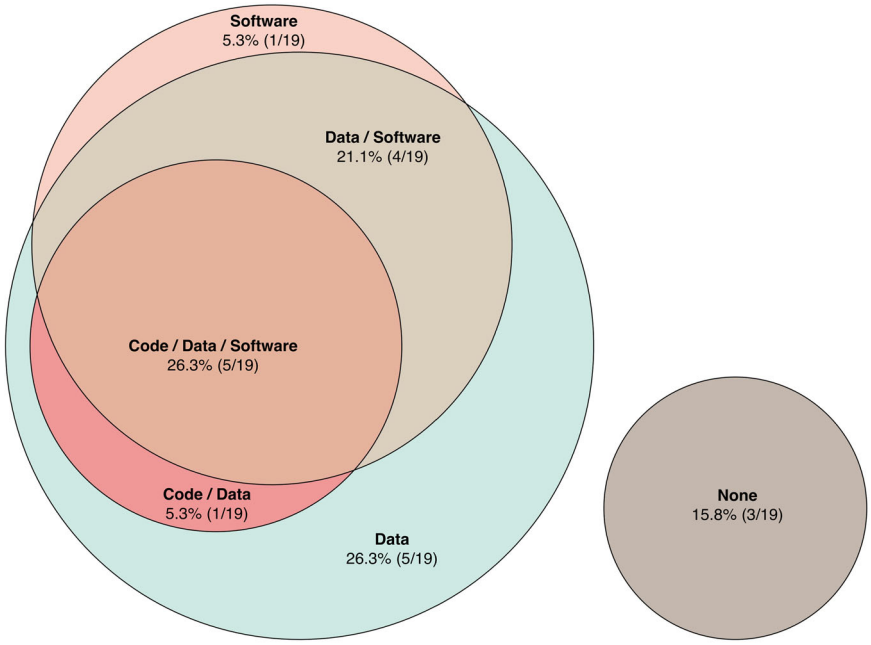

Fig. 2 Open science practices in machine learning for antibiotic discovery. This Euler diagram visualizes public release rates for source code, training or testing data, software, and combinations thereof among publications cited in this review (Table 2). Note that data release criteria for this analysis include both partial and full public availability. This analysis was performed post hoc on studies previously cited in this review.

ML and cancer therapies, ML and cardiovascular drugs, or ML and drugs broadly defined (Fig. 3). Querying for applications of ML to broad drug development serves as a benchmark against which to compare engagement levels in antibiotic-specific applications. Disease group-specific keywords were excluded from the general drug query to prevent double-counting. As cardiovascular disease and cancer are the two leading causes of death in the United States[137], querying for these applications provides relevant public health context for infectious disease applications. Further, a blanket query for AI and ML keywords with no additional qualifiers provides the most macroscopic view of research interest in these predictive methodologies, irrespective of application area. Exact Boolean search phrases can be found in Supplementary Table 1.

Results indicate increasing research interest in all areas over the first two decades of the twenty-first century, with the volume of ML literature focused explicitly on antibiotics and cancer drugs lagging behind broader drug development applications by nearly a decade. Surprisingly, publication counts for cardiovascular drugs and ML remain very low. As the general drug query did not double-count observations from the major disease groups explored, these results may suggest that broad applications have received greater research attention than disease group-specific applications. However, similar trends in cancer- and antibiotic-related publication rates suggest that antibiotics might not be disproportionately neglected. The prevalence of general-interest lines of inquiry might be due to the relative recency of ML for drug discovery, whereby the initial establishment phase lays the groundwork for future specialization. To that end, the significantly higher volume of general drug development applications represents a reservoir of research that is expected to have trickle-down impacts on disease group-specific research over time. Further, the proportion of AI and ML publications that feature applications in general drug discovery, antibiotic discovery, and cancer drug discovery have each increased throughout the twenty-first century. Our analysis also marks 2018 as a watershed moment for the use of ML for antibiotic discovery, coinciding with landmark papers in the field published that year together with preceding software developments.

Over the third decade of the twenty-first century, prospects for ML-facilitated antibiotic discovery will partially hinge on data improvements. As larger data sources become publicly available, new ML questions can be pursued and ongoing questions can be revisited with greater rigor. While expanding public sources of experimental data will be crucial, federated learning across institutions may facilitate empirical dataset expansion without sharing private data, as has been done in other areas of biomedical ML[138]. Increased data sharing from both successful and failed projects in the pharmaceutical industry has also been proposed as a means of accelerating research and development[139]. Existing data can also be further mined for new purposes, as exemplified by resources like the Drug Repurposing Hub[140]. While ML increasingly opens up new regions of chemical space to exploration, the repurposing of non-antibiotic pharmaceuticals could also be a promising avenue for antibiotic discovery[1] that has already benefited from DL methods[16].

A recent review observed greater technical correctness among biomedical ML publications featuring collaborations across computer science, biology, and medicine[129], suggesting that computational antibiotic discovery might similarly benefit from combined expertise. Increased coupling of in silico model testing with in vitro and in vivo validation—and even additional computational methods, e.g., molecular dynamics simulation[141]—will help ensure that published models are robust and yield experimentally actionable predictions. Interdisciplinary collaboration might also facilitate increasingly insightful predictions through biologically informed IML. As a response to the prevalent "black-boxing" of ML models' internal decision-making, IML is an expanding focus in biomedical computation[142] that has been used to elucidate antibiotic MOAs[143]. As firmer terminological and methodological standards alleviate significant confusion surrounding its diverse implementations[144], IML is expected to enable greater human interpretability and causal inference in antibiotic discovery than opaque algorithms generally allow. Expanding interpretability for causal biological insights will surely require both computational creativity and biomedical domain knowledge.

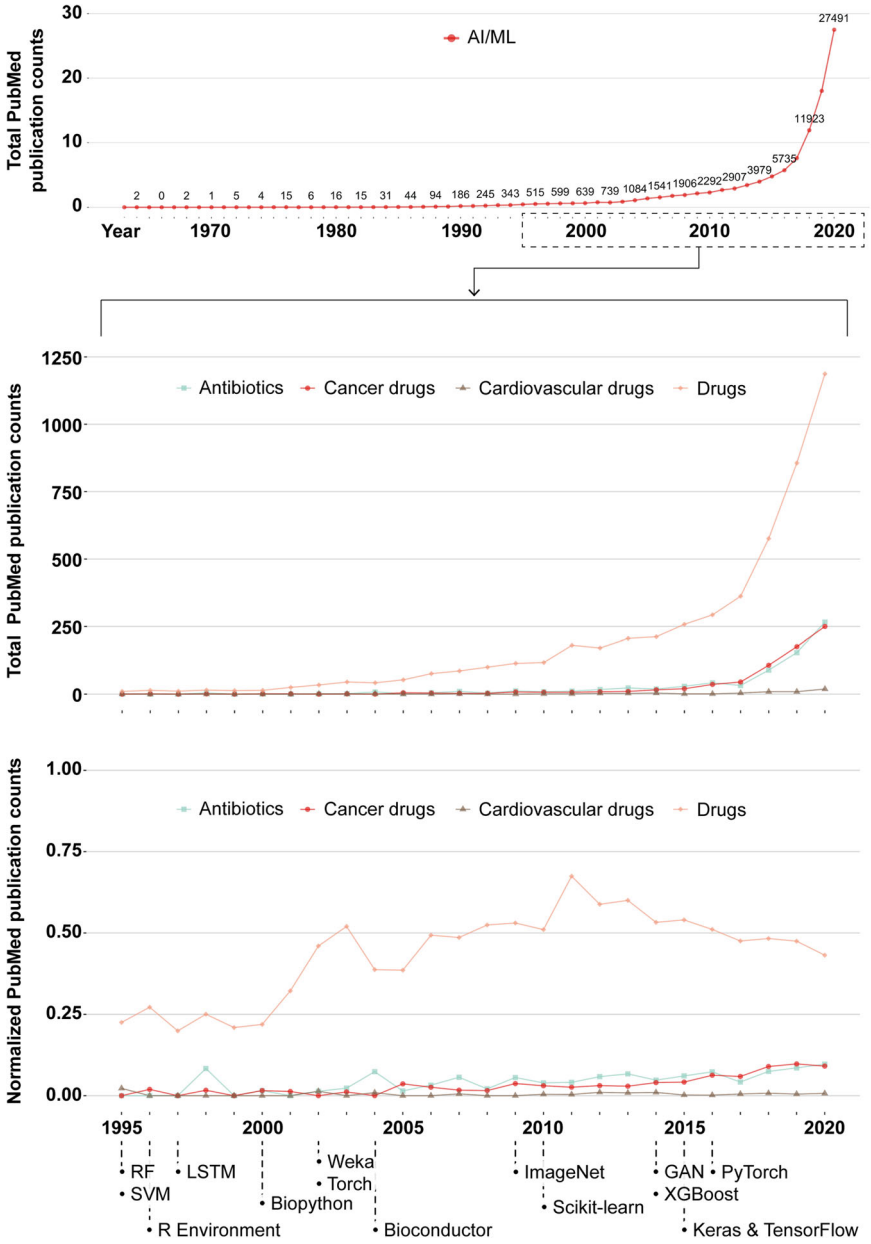

Fig. 3 Machine learning in antibiotic discovery over time. From top to bottom: total PubMed results when querying for AI/ML keywords only, total results when querying for AI/ML and general or disease group-specific drug keywords, and the proportion of general AI/ML publications pertaining to each category of drugs (i.e., total publication counts per drug category scaled by total AI/ML publications per year). Queries sought keywords in titles and abstracts only, with the general drug query excluding keywords contained in the disease group queries to prevent double-counting. Key events in the broader ML community are noted to contextualize trend lines. The relevant literature used to set key dates are as follows: development of SVM[146] and random forest algorithms[147] in 1995; publication of the R language and software environment in 1996[148]; development of LSTM in 1997[149]; development of the Biopython package in 2000[150]; release of the Java interface for Weka in 2002[151]; publication of the Torch library in 2002[152]; release of Bioconductor in 2004[153]; the publication of ImageNet in 2009[154]; the initial release of Scikit-learn in 2010[155]; the initial release of XGBoost[156] and development of GANs[102] in 2014; development of Keras[157] and TensorFlow[158] in 2015; and the initial release of PyTorch in 2016[159]. Exact Boolean searches in PubMed can be found in Supplementary Table 1.

Additional new avenues for ML-facilitated antibiotic discovery are expected to trickle in from algorithmic theory, robotic AI, and adjacent computational domains. While this review has focused on ML rather than embodied AI, recent attempts to deploy intelligent robots in chemical experimentation[145] may indicate the utility of ML-guided autonomous robotics in future antibiotic discovery. Creative integration of diverse lessons from NLP, computer vision, generative DL, computer-aided drug design, and other flourishing areas in ML research will play important roles in accelerating the urgent task of novel antibiotic discovery.

Data availability

CSV files containing the raw PubMed data outputs visualized in Fig. 3 are available in Supplementary Data 1. A README file containing resource metadata is also provided.

Received: 18 February 2021; Accepted: 16 July 2021;
Published online: 09 September 2021

References

1. De Oliveira, D. M. et al. Antimicrobial resistance in ESKAPE pathogens. *Clin. Microbiol. Rev.* **33**, 1–49 (2020).
2. CDC. *Antibiotic Resistance Threats in the United States, 2019*. Technical Report (US Department of Health and Human Services, CDC, 2019).
3. Chng, K. R. et al. Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nat. Ecol. Evol.* https://doi.org/10.1038/s41559-020-1236-0 (2020).
4. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
5. Lepore, C., Silver, L., Theuretzbacher, U., Thomas, J. & Visi, D. The small-molecule antibiotics pipeline: 2014-2018. *Nat. Rev. Drug Discov.* **18**, 739–739 (2019).
6. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
7. Durrant, J. D. & Amaro, R. E. Machine-learning techniques applied to antibacterial drug discovery. *Chem. Biol. Drug Des.* **85**, 14–21 (2015).
8. de la Fuente-Nunez, C. Toward autonomous antibiotic discovery. *mSystems* **4**, 10–14 (2019).
9. Torres, M. D. T. & de la Fuente-Nunez, C. Toward computer-made artificial antibiotics. *Curr. Opin. Microbiol.* **51**, 30–38 (2019).
10. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
11. Porto, W. F. et al. In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun.* **9**, 1490 (2018).
12. Torres, P. H. M., Sodero, A. C. R., Jofily, P. & Silva-Jr, F. P. Key topics in molecular docking for drug design. *Int. J. Mol. Sci.* **20**, 4574 (2019).
13. Adeshina, Y. O., Deeds, E. J. & Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl Acad. Sci. USA* **117**, 18477–18488 (2020).
14. Pereira, J. C., Caffarena, E. R. & dos Santos, C. N. Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* **56**, 2495–2506 (2016).
15. Li, H., Sze, K. H., Lu, G. & Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev.* **10**, 1–20 (2020).
16. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
17. E. Greber, K. & Dawgul, M. Antimicrobial peptides under clinical trials. *Curr. Top. Med. Chem.* **17**, 620–628 (2017).
18. Magana, M. et al. The value of antimicrobial peptides in the age of resistance. *Lancet Infect. Dis.* **20**, e216–e230 (2020).
19. Cherkasov, A. et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
20. Kawashima, S. AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374–374 (2000).
21. Brüstle, M. et al. Descriptors, physical properties, and drug-likeness. *J. Med. Chem.* **45**, 3345–3355 (2002).
22. Speck-Planche, A. Multicellular Target QSAR Model for simultaneous prediction and design of anti-pancreatic cancer agents. *ACS Omega* **4**, 3122–3132 (2019).
23. Prado-Prado, F. et al. 3D MI-DRAGON: new model for the reconstruction of US FDA drug-target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Curr. Top. Med. Chem.* **12**, 1843–1865 (2012).
24. van Westen, G. J. et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *J. Cheminformatics* **5**, 41 (2013).
25. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
26. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A. & Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Systems* 1–10, https://doi.org/10.1016/j.cels.2020.08.016 (2020).
27. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
28. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J.-L. Drug analogs from fragment-based long short-term memory generative neural networks. *J. Chem. Inf. Model.* **59**, 1347–1356 (2019).
29. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, 1–15 (2018).
30. Müller, A. T., Hiss, J. A. & Schneider, G. Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model.* **58**, 472–479 (2018).
31. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* https://doi.org/10.1038/s41592-019-0598-1 (2019).
32. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* http://www.nature.com/articles/s41573-019-0050-3 (2019).
33. Mansbach, R. A. et al. Machine learning algorithm identifies an antibiotic vocabulary for permeating Gram-negative bacteria. *J. Chem. Inf. Model.* **60**, 2838–2847 (2020).
34. Sagi, O. & Rokach, L. Ensemble learning: a survey. *Wiley Interdiscip. Rev.* **8**, 1–18 (2018).
35. Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 1–15 (Springer, 2000).
36. Manavalan, B., Shin, T. H., Kim, M. O. & Lee, G. PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* **9**, 1783 (2018).
37. Zhang, W. et al. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* **173**, 979–987 (2016).
38. Lee, E. Y., Fulan, B. M., Wong, G. C. L. & Ferguson, A. L. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc. Natl Acad. Sci. USA* **113**, 13588–13593 (2016).
39. Lee, M. W., Lee, E. Y., Ferguson, A. L. & Wong, G. C. Machine learning antimicrobial peptide sequences: some surprising variations on the theme of amphiphilic assembly. *Curr. Opin. Colloid Interface Sci.* **38**, 204–213 (2018).
40. Fjell, C. D. et al. Identification of novel antibacterial peptides by chemoinformatics and machine learning †. *J. Med. Chem.* **52**, 2006–2015 (2009).
41. Yan, J. et al. Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. *Mol. Ther. Nucleic Acids* **20**, 882–894 (2020).
42. Nguyen, M. et al. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumonia*. *Sci. Rep.* **8**, 1–11 (2018).
43. Nguyen, M. et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. *J. Clin. Microbiol.* **57**, 1–15 (2019).
44. Nagarajan, D. et al. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J. Biol. Chem.* **293**, 3492–3509 (2018).
45. Dean, S. N. & Walper, S. A. Variational autoencoder for generation of antimicrobial peptides. *ACS Omega* **5**, 20746–20754 (2020).
46. Fjell, C. D., Hiss, J. A., Hancock, R. E. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11**, 37–51 (2012).
47. Vishnepolsky, B. et al. Predictive model of linear antimicrobial peptides active against Gram-negative bacteria. *J. Chem. Inf. Model.* **58**, 1141–1151 (2018).
48. Vishnepolsky, B. et al. De novo design and in vitro testing of antimicrobial peptides against gram-negative bacteria. *Pharmaceuticals* **12**, 82 (2019).
49. Yoshida, M. et al. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* **4**, 533–543 (2018).
50. Zoffmann, S. et al. Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **9**, 5013 (2019).
51. Lipinski, C. A. Rule of five in 2015 and beyond: target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. *Adv. Drug Deliv. Rev.* **101**, 34–41 (2016).
52. Jia, C. Y., Li, J. Y., Hao, G. F. & Yang, G. F. A drug-likeness toolbox facilitates ADMET study in drug discovery. *Drug Discov. Today* **25**, 248–258 (2020).
53. D'Souza, S., Prema, K. V. & Balaji, S. Machine learning models for drug-target interactions: current knowledge and future directions. *Drug Discov. Today* **25**, 748–756 (2020).
54. Timmons, P. B. & Hewage, C. M. OPEN HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Scientific Rep.* 1–18, https://doi.org/10.1038/s41598-020-67701-3 (2020).
55. Cruz-Monteagudo, M., Borges, F. & Cordeiro, M. N. D. S. Jointly handling potency and toxicity of antimicrobial peptidomimetics by simple rules from desirability theory and chemoinformatics. *J. Chem. Inf. Model.* **51**, 3060–3077 (2011).
56. Plisson, F., Sánchez, O. R. & Hernández, C. M. Machine learning-guided discovery and design of non-hemolytic peptides. *Scientific Rep.* 1–19, https://doi.org/10.1038/s41598-020-73644-6 (2020).
57. Zheng, S. et al. Quantitative prediction of hemolytic toxicity for small molecules and their potential hemolytic fragments by machine learning and recursive fragmentation methods. *J. Chem. Inf. Model.* **60**, 3231–3245 (2020).
58. Zheng, S. et al. Computational prediction of a new ADMET endpoint for small molecules: anticommensal effect on human gut microbiota. *J. Chem. Inf. Model.* **59**, 1215–1220 (2019).

59. Webel, H. E. et al. Revealing cytotoxic substructures in molecules using deep learning. *J. Comput. Aided Mol. Des.* **34**, 731–746 (2020).
60. Gao, M., Igata, H., Takeuchi, A., Sato, K. & Ikegaya, Y. Machine learning-based prediction of adverse drug effects: an example of seizure-inducing compounds. *J. Pharmacol. Sci.* **133**, 70–78 (2017).
61. Khurana, S. et al. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).
62. Han, X., Zhang, L., Zhou, K. & Wang, X. ProGAN: protein solubility generative adversarial nets for data augmentation in DNN framework. *Comput. Chem. Eng.* **131**, 106533 (2019).
63. Rawi, R. et al. PaRSnIP: Sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34**, 1092–1098 (2018).
64. Smialowski, P., Doose, G., Torkler, P., Kaufmann, S. & Frishman, D. PROSO II - a new method for protein solubility prediction. *FEBS J.* **279**, 2192–2200 (2012).
65. Han, X., Wang, X., Zhou, K. & Valencia, A. Develop machine learning-based regression predictive models for engineering protein solubility. *Bioinformatics* **35**, 4640–4646 (2019).
66. Hou, Q., Kwasigroch, J. M., Rooman, M. & Pucci, F. SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* **36**, 1445–1452 (2020).
67. Torres, M. D., Sothiselvam, S., Lu, T. K. & de la Fuente-Nunez, C. Peptide design principles for antimicrobial applications. *J. Mol. Biol.* **431**, 3547–3567 (2019).
68. Der Torossian Torres, M. & De La Fuente-Nunez, C. Reprogramming biological peptides to combat infectious diseases. *Chem. Commun.* **55**, 15020–15032 (2019).
69. Radchenko, T., Fontaine, F., Morettoni, L. & Zamora, I. Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery. *PLoS ONE* **14**, 1–20 (2019).
70. Wang, P. et al. Multi-label learning for predicting the activities of antimicrobial peptides. *Sci. Rep.* **7**, 1–11 (2017).
71. Wee, L. J., Tan, T. W. & Ranganathan, S. CASVM: Web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics* **23**, 3241–3243 (2007).
72. Piippo, M., Lietzén, N., Nevalainen, O. S., Salmi, J. & Nyman, T. A. Pripper: prediction of caspase cleavage sites from whole proteomes. *BMC Bioinform.* **11**, 1–9 (2010).
73. Song, J. et al. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS ONE* **7**, e50300 (2012).
74. Song, J. et al. IProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.* **20**, 638–658 (2019).
75. Li, F. et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* **36**, 1057–1065 (2020).
76. Li, F. et al. Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. *Genom. Proteom. Bioinform.* **18**, 52–64 (2020).
77. Song, J. et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* **34**, 684–687 (2018).
78. Li, X. et al. DeepChemStable: chemical stability prediction with an attention-based graph convolution network. *J. Chem. Inf. Model.* **59**, 1044–1049 (2019).
79. Liu, Z. et al. ChemStable: a web server for rule-embedded naïve Bayesian learning approach to predict compound stability. *J. Comput. Aided Mol. Des.* **28**, 941–950 (2014).
80. Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**, 881–890 (2007).
81. Macesic, N., Polubriaginof, F. & Tatonetti, N. P. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. *Curr. Opin. Infect. Dis.* **30**, 511–517 (2017).
82. Hicks, A. L. et al. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLoS Comput. Biol.* **15**, e1007349 (2019).
83. Pizzo, E., Cafaro, V., Di Donato, A. & Notomista, E. Cryptic antimicrobial peptides: identification methods and current knowledge of their immunomodulatory properties. *Curr. Pharm. Des.* **24**, 1054–1066 (2018).
84. de Oliveira Costa, B. & Franco, O. L. Cryptic host defense peptides: multifaceted activity and prospects for medicinal chemistry. *Curr. Top. Med. Chem.* **20**, 1274–1290 (2020).
85. Lázár, V. et al. Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nat. Microbiol.* **3**, 718–731 (2018).
86. Hyun, J. C., Kavvas, E. S., Monk, J. M. & Palsson, B. O. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput. Biol.* **16**, 1–24 (2020).
87. Her, H. L. & Wu, Y. W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics* **34**, i89–i95 (2018).
88. Moradigaravand, D. et al. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput. Biol.* **14**, 1–17 (2018).
89. Khaledi, A. et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol. Med.* **12**, 1–19 (2020).
90. Yang, Y. et al. DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis*. *Bioinformatics* **35**, 3240–3249 (2019).
91. Yang, Y. et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34**, 1666–1671 (2018).
92. Deelder, W. et al. Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data. *Front. Genet.* **10**, 1–9 (2019).
93. Davis, J. J. et al. Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.* **6**, 1–12 (2016).
94. Chowdhury, A. S., Khaledian, E. & Broschat, S. L. Capreomycin resistance prediction in two species of Mycobacterium using a stacked ensemble method. *J. Appl. Microbiol.* **127**, 1656–1664 (2019).
95. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 1–15 (2018).
96. Kim, J. et al. VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Computl. Biol.* **16**, 1–17 (2020).
97. Pandey, D., Kumari, B., Singhal, N. & Kumar, M. BacEffluxPred: a two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins. *Sci. Rep.* **10**, 1–9 (2020).
98. Rahman, S. F., Olm, M. R., Morowitz, M. J. & Banfield, J. F. Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *mSystems* **3**, 1–12 (2018).
99. Ruppé, E. et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat. Microbiol.* **4**, 112–123 (2019).
100. Kavvas, E. S. et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, https://doi.org/10.1038/s41467-018-06634-y (2018).
101. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.* **11**, 1–11 (2020).
102. Goodfellow, I. et al. Generative adversarial nets. In *Proc. 2014 Advances in Neural Information Processing Systems* 2672–2680 (2014).
103. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In *Second International Conference on Learning Representations, ICLR 2014—Conference Track Proceedings*, 1–14 (2014).
104. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
105. Dan, Y. et al. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* **6**, 1–7 (2020).
106. Pierce, N. A. & Winfree, E. Protein design is NP-hard. *Protein Eng.* **15**, 779–782 (2002).
107. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Adv. Neural Inform. Process. Syst.* **32**, 15820–15831 (2019).
108. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. DruGAN: an Advanced Generative Adversarial Autoencoder Model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **14**, 3098–3104 (2017).
109. Putin, E. et al. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **58**, 1194–1204 (2018).
110. Putin, E. et al. Adversarial threshold neural computer for molecular de novo design. *Mol. Pharm.* **15**, 4386–4397 (2018).
111. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
112. Prykhodko, O. et al. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminformatics* **11**, 1–13 (2019).
113. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminformatics* **9**, 1–14 (2017).
114. Kotsias, P.-C. et al. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2**, 254–265 (2020).
115. Arús-Pous, J. et al. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminformatics* **11**, 1–13 (2019).

116. Arús-Pous, J. et al. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminformatics* **12**, 1–18 (2020).
117. Grisoni, F. et al. Designing anticancer peptides by constructive machine learning. *ChemMedChem* **13**, 1300–1302 (2018).
118. Grisoni, F. et al. De novo design of anticancer peptides by ensemble artificial neural networks. *J. Mol. Model.* **25**, 1–10 (2019).
119. Tucs, A. et al. Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks. *ACS Omega* **5**, 22847–22851 (2020).
120. Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
121. Getahun, H., Smith, I., Trivedi, K., Paulina, S. & Balkhy, H. H. Tackling antimicrobial resistance in the COVID-19 pandemic. *Bull. World Health Organ.* **98**, 441–508 (2020).
122. Cutler, D. M. & Summers, L. H. The COVID-19 pandemic and the $16 trillion virus. *JAMA* **324**, 1495–1496 (2020).
123. Karaca-Mandic, P., Georgiou, A. & Sen, S. Assessment of COVID-19 hospitalizations by race/rthnicity in 12 states. *JAMA Intern. Med.* **181**, 131–134 (2020).
124. Homolak, J., Kodvanj, I. & Virag, D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics* **124**, 2687–2701 (2020).
125. Schiltz, M. Science without publication paywalls: cOAlition S for the realisation of full and immediate open access. *PLoS Med.* **15**, 2018–2021 (2018).
126. Haibe-Kains, B. et al. Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).
127. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
128. Beam, A. L., Manrai, A. K. & Ghassemi, M. Challenges to the reproducibility of machine learning models in health care. *JAMA* **323**, 305–306 (2020).
129. Littmann, M. et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat. Mach. Intell.* **2**, 18–24 (2020).
130. McDermott, M. B. et al. Reproducibility in machine learning for health research: still a ways to go. *Sci. Transl. Med.* **13**, eabb1655 (2021).
131. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
132. Fujihashi, M. et al. An unprecedented twist to ODCase catalytic activity. *J. Am. Chem. Soc.* **127**, 15048–15050 (2005).
133. Brainard, J. California universities and Elsevier make up, ink big open-access deal. *Science* https://www.sciencemag.org/news/2021/03/california-universities-and-elsevier-make-ink-big-open-access-deal (2021).
134. Brainard, J. A new mandate highlights costs, benefits of making all scientific articles free to read. *Science* https://www.sciencemag.org/news/2021/01/new-mandate-highlights-costs-benefits-making-all-scientific-articles-free-read (2021).
135. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
136. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence and Medical Imaging (CLAIM). *Radiol. Artif. Intell.* **2**, e200029 (2020).
137. Kochanek, K. D., Xu, J. & Arias, E. Mortality in the United States, 2019. *Centers for Disease Control and Prevention NCHS Data Brief*, Vol. 395 (National Center for Health Statistics, 2020).
138. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 1–12 (2020).
139. Kim, W., Krause, K., Zimmerman, Z. & Outterson, K. Improving data sharing to increase the efficiency of antibiotic R&D. *Nat. Rev. Drug Discov.* https://www.nature.com/articles/d41573-020-00185-y (2020).
140. Corsello, S. M. et al. *Inf. Resour.* **23**, 405–408 (2017).
141. Melo, M. C., Bernardi, R. C., De La Fuente-Nunez, C. & Luthey-Schulten, Z. Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *J. Chem. Phys.* **153**, https://doi.org/10.1063/5.0018980 (2020).
142. Yu, M. K. et al. Visible machine learning for biomedicine. *Cell* **173**, 1562–1565 (2018).
143. Yang, J. H. et al. A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell* **177**, 1649–1661 (2019).
144. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
145. Burger, B. et al. A mobile robotic researcher. *Nature* **583**, https://doi.org/10.1038/s41586-020-2442-2 (2020).
146. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
147. Ho, T. K. Random decision forests. In *Proc. 3rd International Conference on Document Analysis and Recognition*, Vol. 1, 278–282 (IEEE Comput. Soc. Press, 1995).
148. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
149. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
150. Chapman, B. & Chang, J. Biopython: Python tools for computational biology. *ACM SIGBIO Newsl.* **20**, 15–19 (2000).
151. Witten, I. H. & Frank, E. Data mining: practical machine learning tools and techniques with Java implementations. *ACM Sigmod Rec.* **31**, 76–77 (2002).
152. Collobert, R., Bengio, S. & Marithoz, J. *Torch: A Modular Machine Learning Software Library*. Technical Report 02-46 (IDIAP, 2002).
153. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
154. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *IEEE Computer Vision and Pattern Recognition*, 248–255 (2009).
155. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
156. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, New York, 2016).
157. Chollet, F. Keras: deep learning library for theano and tensorflow. https://keras.io/k (2015).
158. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283 ({USENIX} Association, Savannah, 2016).
159. Paszke, A. et al. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff* (2017).
160. Smith, R. D. et al. Updates to binding MOAD (Mother of All Databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.* **431**, 2423–2433 (2019).
161. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
162. Chang, A. et al. BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.* **43**, D439–D446 (2015).
163. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
164. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
165. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
166. Burley, S. K. et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
167. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
168. Szklarczyk, D. et al. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380–D384 (2016).
169. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2016).
170. Hecker, N. et al. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.* **40**, 1113–1117 (2012).
171. Wang, Y. et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **48**, D1031–D1041 (2020).
172. Bateman, A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
173. Sterling, T. & Irwin, J. J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
174. Lee, H. T. et al. A large-scale structural classification of antimicrobial peptides. *Biomed. Res. Int.* **2015**, 475062 (2015).
175. Ramos-Martín, F., Annaval, T., Buchoux, S., Sarazin, C. & D'Amelio, N. Adaptable: a comprehensive web platform of antimicrobial peptides tailored to the user's research. *Life Sci. Alliance* **2**, e201900512 (2019).
176. Waghu, F. H., Barai, R. S., Gurung, P. & Idicula-Thomas, S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* **44**, D1094–D1097 (2016).
177. Kang, X. et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data* **6**, 1–10 (2019).
178. Pirtskhalava, M. et al. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2020).
179. Jhong, J. H. et al. DbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Res.* **47**, D285–D297 (2019).
180. Doster, E. et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.* **48**, D561–D569 (2020).
181. Davis, J. J. et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).

182. Urán Landaburu, L. et al. TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. *Nucleic Acids Res.* **48**, D992–D1005 (2020).
183. Chaudhary, K. et al. A web server and mobile app for computing hemolytic potency of peptides. *Sci. Rep.* **6**, 1–13 (2016).

Acknowledgements

Cesar de la Fuente-Nunez holds a Presidential Professorship at the University of Pennsylvania, is a recipient of the Langer Prize by the AIChE Foundation and acknowledges funding from the Institute for Diabetes, Obesity, and Metabolism, the Penn Mental Health AIDS Research Center of the University of Pennsylvania, the Nemirovsky Prize, the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania, the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138201, and the Defense Threat Reduction Agency (DTRA; HDTRA11810041 and HDTRA1-21-1-0014). J.R.M.A.M. acknowledges support from the University of Pennsylvania GAPSA-Provost Fellowship for Interdisciplinary Innovation and the Open Knowledge Foundation Frictionless Data for Reproducible Research Fellowship, funded by the Alfred P. Sloan Foundation.

Author contributions

M.C.R.M. and J.R.M.A.M. wrote the original draft. M.C.R.M., J.R.M.A.M., and C.d.l.F.-N. reviewed the final manuscript.

Competing interests

C.d.l.F.-N. is an Editorial Board Member for *Communications Biology*, but was not involved in the editorial review of, nor the decision to publish, this article. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02586-0.

Correspondence and requests for materials should be addressed to Cesar de la Fuente-Nunez.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Luke R. Grinham.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.