DOI: 10.1111/nyas.14930

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

REVIEW

Antibiotic discovery in the artificial intelligence era

Telmah Lluka | Jonathan M. Stokes

Department of Biochemistry and Biomedical Sciences, Michael G. DeGroote Institute for Infectious Disease Research, David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada

Correspondence
Jonathan M. Stokes, Department of Biochemistry and Biomedical Sciences, McMaster University, 1280 Main St W, Hamilton, ON L8S 4L8, Canada.
Email: stokesjm@mcmaster.ca

Funding information
Weston Family Foundation; David Braley Centre for Antibiotic Discovery; Natural Sciences and Engineering Research Council of Canada

Abstract

As the global burden of antibiotic resistance continues to grow, creative approaches to antibiotic discovery are needed to accelerate the development of novel medicines. A rapidly progressing computational revolution—artificial intelligence—offers an optimistic path forward due to its ability to alleviate bottlenecks in the antibiotic discovery pipeline. In this review, we discuss how advancements in artificial intelligence are reinvigorating the adoption of past antibiotic discovery models—namely natural product exploration and small molecule screening. We then explore the application of contemporary machine learning approaches to emerging areas of antibiotic discovery, including antibacterial systems biology, drug combination development, antimicrobial peptide discovery, and mechanism of action prediction. Lastly, we propose a call to action for open access of high-quality screening datasets and interdisciplinary collaboration to accelerate the rate at which machine learning models can be trained and new antibiotic drugs can be developed.

KEYWORDS
antibiotics, drug discovery, machine learning

INTRODUCTION

In 2019, antibiotic-resistant infections were estimated to have contributed to 4.95 million deaths worldwide,[1] a number that is expected to increase to 10 million deaths per year by 2050 without immediate productivity in discovering new antibiotics.[2] Worryingly, the global dissemination of antibiotic resistance is exacerbated by an alarmingly lean antibiotic development pipeline. Of the 45 antibacterial compounds in clinical development as of November 2021, only six of them are both (1) active against the World Health Organization's (WHO) priority pathogens and (2) considered innovative as defined by the WHO's innovation criteria—no known cross-resistance, novel target, a novel mode of action, and/or novel class.[3] A major contributor to this innovation gap is a lack of economic incentives to motivate antibiotic research and development by large pharmaceutical companies.[4] As a result, the responsibility has largely been placed on academic laboratories and small biotechnology enterprises to fulfill the discovery and preclinical stages of the antibiotic development pipeline—including hit identification, mechanism of action (MOA) elucidation, and hit-to-lead medicinal chemistry efforts. Unfortunately, these early stages are often costly, laborious, and characterized by a high attrition rate. For instance, from 2011 to 2020, there was only an ~16% FDA approval rate for antibacterial compounds in phase I clinical trials.[5] Moreover, the cost of developing an antibiotic is an estimated $1.5 billion;[6] however, this expense is rarely recuperated. In 2018, the cephalosporin antibiotic ceftaroline was the highest-selling antibiotic in the United States, generating $138 million in revenue. In contrast, the most profitable nonantimicrobial drug earned $13.68 billion in the same year.[7] The unsustainability of the current antibiotic drug discovery paradigm is especially evident in the failure of the biotechnology firm Achaogen, which filed for bankruptcy in 2019 despite the FDA approval of their antibiotic, plazomicin—an aminoglycoside developed for the treatment of multidrug-resistant Enterobacteriaceae urinary tract infections.[8]

Despite continuous and rapid advancements in genomics, laboratory automation, and synthetic chemistry methods, we have failed to translate these into novel antibiotics for clinical use[9]—a phenomenon appropriately termed the "knowledge paradox."[10] We posit that this is, in part, because we are underexploiting the dense, multidimensional datasets[11] that are commonly generated during the large screening campaigns associated with modern drug discovery programs.[12]

Such datasets are typically manually analyzed to shortlist hit chemicals for downstream analysis, but their size and complexity makes it difficult to observe and rationalize the latent relationships between bacterial physiological and chemical perturbation.[13] For reference, modern techniques to profile antibiotic stress allow for the collection of chemical–genetic interactions and biomass measurements across hundreds of time points.[14] Indeed, one chemogenomic study[15] collected nearly 20 million observations for 15 antibiotics against a library of ∼4000 nonessential single-gene deletion *Escherichia coli* mutants.[16] Datasets of this scale are challenging to interpret, requiring sophisticated approaches to maximize the utility of such large-scale data and translate these into new antibiotics.

A promising approach is the application of machine learning (ML) techniques to antibiotic discovery. ML methods have rapidly grown in popularity within many scientific disciplines—other fields of drug discovery included—where they have shown utility in relieving bottlenecks within the preclinical and clinical development pipelines.[17–19] Indeed, ML is emerging as a powerful and cost-effective tool that pharmaceutical companies have recently adopted in their own drug discovery efforts. For example, Roche and Genentech are collaborating with Recursion Pharmaceuticals to help accelerate their research into neuroscience and oncology therapies. Novartis has partnered with Microsoft to establish an artificial intelligence innovation lab with the goal of accelerating its molecular design and personalized therapy programs. Sanofi has entered into a partnership worth up to $5.2 billion with Exscientia to develop new drugs for oncology and immunology using artificial intelligence, among numerous other examples.[20] However, ML approaches have yet to be broadly implemented in antibiotics research, which is detrimental to the field since these methods are well suited to accelerate the antibiotic discovery pipeline[21]—for example, by enabling the rapid exploration of vast chemical and sequence spaces[22]—thereby increasing the likelihood of discovering novel structural and functional classes of antibiotics, while decreasing associated costs.

Despite the promise of ML in antibiotic discovery, we must remember that ML is not a panacea. Antibiotic drug discovery is a challenging endeavor that involves the complex physiologies of both humans and bacteria; it requires a molecule to be optimized for multiple properties, such as low cytotoxicity, favorable pharmacokinetics/pharmacodynamics (PK/PD), and high potency against a pathogen of interest. As such, ML approaches require large, high-quality wet lab datasets on which to train[23] and sufficient computational fluency to implement the correct algorithms in the correct manner. It is difficult to ignore the similarities between ML in drug discovery and the genomics era of antibiotic discovery, the latter of which saw the introduction of high-throughput screening methods and recombinant DNA technology. This period failed to develop viable clinical antibiotics[24] in part due to diminishing investment in the later stages of drug development,[24,25] but also due to the reductionist nature of the empirical methods used to identify hit compounds. Similarly, ML approaches should not be used without deep consideration; the success of ML methods relies on their appropriate implementation to enhance—not necessarily replace—current methods.

In this review, we explore the evolution of dominant antibiotic discovery approaches since the beginning of the 20th century, highlighting how technological advancements were instrumental to our successes in novel antibiotic discovery, and how ML approaches may augment each antibiotic discovery technology to help us more rapidly discover novel drugs (Figure 1). We will first describe how ML approaches can be used to augment natural product (NP) discovery. Next, we will discuss the application of ML in expanding the chemical search space in novel small-molecule antibiotic discovery. Third, we will discuss the various emerging applications of ML in unconventional antibiotic discovery and development. Lastly, we will conclude the paper with a call to action for more open data sharing and increased multidisciplinary collaboration toward the development of high-quality datasets and robust ML models that are shared among the global community of antibiotic discoverers.

ML CONCEPTS IN A NUTSHELL

ML is an algorithmic approach to building predictive models that are representative of a given dataset. ML models refine themselves in an automated manner by measuring their own predictive performance against a training dataset and optimizing their parameters accordingly. This is the process of "learning." ML approaches are being developed for a wide variety of tasks in drug discovery, including molecular property prediction,[26] small molecule design,[27] MOA elucidation,[28] image-based profiling,[29] and target identification.[30] The appropriate use of ML approaches requires a detailed understanding of the specific goals of a prediction task, as well as the datasets on which to train the ML models. Indeed, these fundamental considerations help determine how the data will be processed, which ML model architectures should be considered, and how to evaluate model performance for downstream real-world applications. In this section, we provide a brief overview of these three considerations. We direct readers to additional thorough guides for further studying.[31–33]

Data preprocessing

The performance of ML models can be continuously improved through providing larger quantities of training data, as well as optimizing the parameters of a chosen model architecture. Of course, extreme care must be exercised in collecting a large quantity of high-quality data to ensure the model can make accurate predictions when applied to real-world tasks. In the context of chemical screens, for example, raw screening data must be encoded as a vector representation for input into the model (Figure 2A). The model will then learn features of the input data (chemical structure or amino acid sequence) in the context of the measurement on which it is being trained (growth inhibition of a bacterium of interest). In practice, compound bioactivity (e.g., growth inhibition) is often binarized such that all molecules with bioactivity above a certain threshold are labeled as 1 and those with bioactivity below the threshold as 0 (see Ref. 34), although regression models can

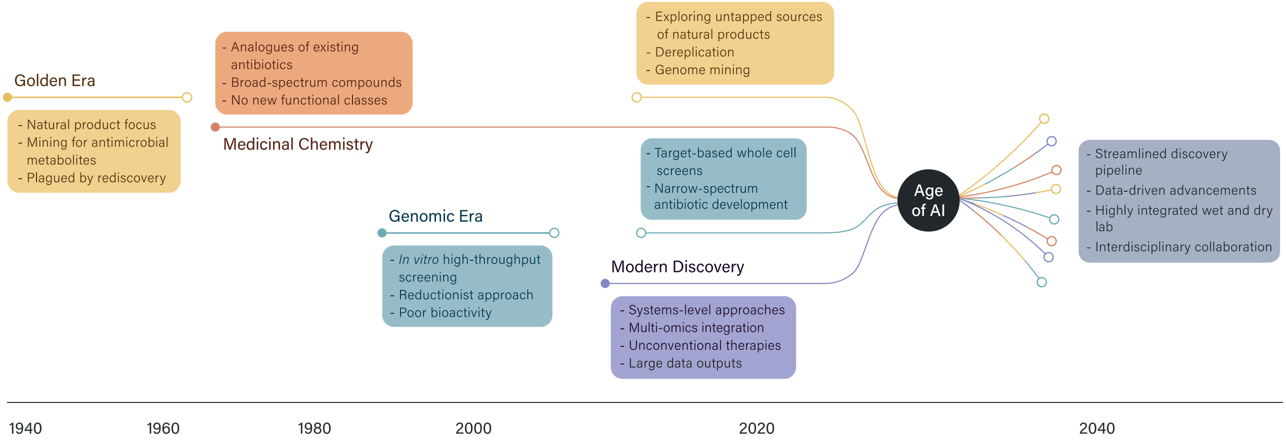
Golden Era
- Natural product focus
- Mining for antimicrobial metabolites
- Plagued by rediscovery
- Analogues of existing antibiotics
- Broad-spectrum compounds
- No new functional classes
Medicinal Chemistry
Genomic Era
- *In vitro* high-throughput screening
- Reductionist approach
- Poor bioactivity
- Exploring untapped sources of natural products
- Dereplication
- Genome mining
- Target-based whole cell screens
- Narrow-spectrum antibiotic development
Modern Discovery
- Systems-level approaches
- Multi-omics integration
- Unconventional therapies
- Large data outputs
Age of AI
- Streamlined discovery pipeline
- Data-driven advancements
- Highly integrated wet and dry lab
- Interdisciplinary collaboration
1940
1960
1980
2000
2020
2040

FIGURE 1 Timeline of major antibiotic discovery approaches. The golden era was a successful period defined by whole-cell screens of secondary metabolites produced by soil-dwelling microbes; it was largely abandoned due to the rediscovery problem. The medicinal chemistry era focused on modifying existing antibiotics to optimize their medicinal and antibacterial properties. In the genomic era, a focus on target-based *in vitro* screens failed to identify any new clinical antibiotics. Modern approaches to antibiotic discovery include an array of unique approaches, with this diversification reflecting a newfound appreciation for antibiotic function in the context of the whole cell and the host infection environment. Within the next decade, end-to-end integration of ML algorithms into existing antibiotic discovery techniques may have the potential to accelerate antibiotic development.

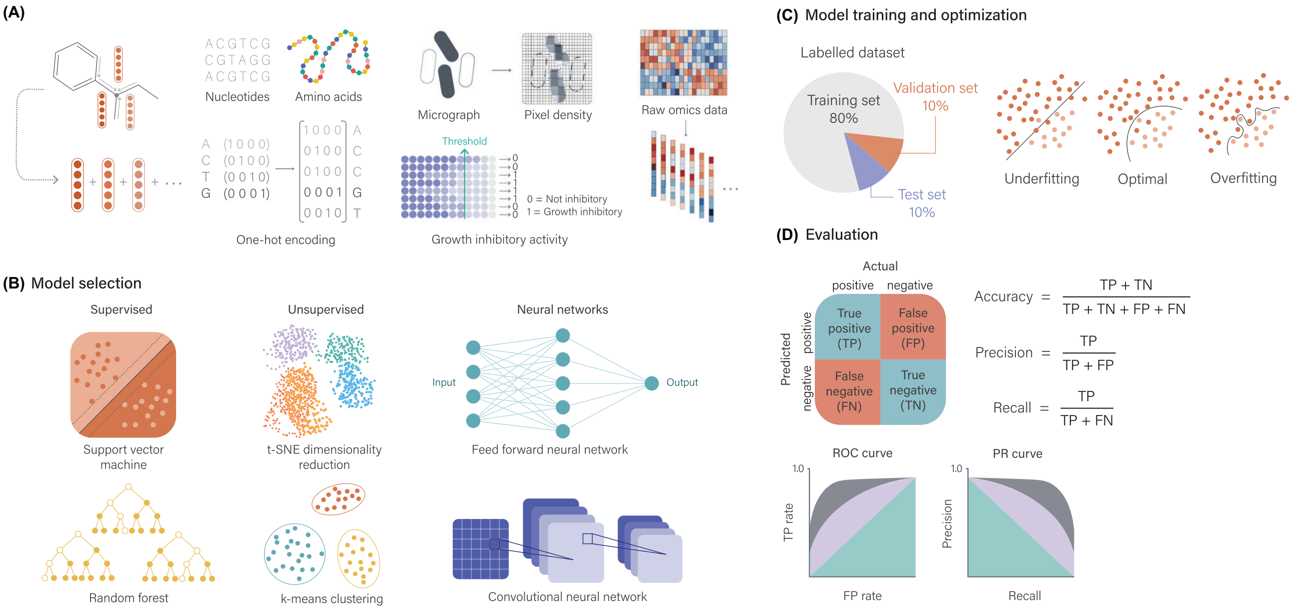

FIGURE 2 Overview of machine learning concepts. (A) Vectorized representations of data relevant to antibiotic discovery research. Molecular structures can be converted to vector representations using graph neural networks (left). One-hot encoding maps each character of a sequence (amino acids or nucleotides) onto a value of 0 or 1 (middle). Feature profiles can be constructed from a variety of different readouts, such as raw pixel intensity from bacterial micrographs or from holistic omics analyses (right). Desired chemical properties can be binarized. Here, growth inhibition is labeled as 1 and no growth inhibition is labeled as 0. (B) Commonly used supervised machine learning models include support vector machines (SVMs) and random forest models (RFs) (left). Commonly used unsupervised learning models include *k*-means clustering and t-distributed stochastic neighbor embedding (t-SNE) (middle). DL models are a subclass of ML that use neural networks to learn. Depicted are two commonly used models, a feed-forward neural network and a convolutional neural network (right). Convolutional neural networks (CNNs) are unique in that they can be used with images; images are first analyzed as raw pixel intensity, then transformed into a vector after a series of alternating convolutional and pooling steps. (C) A labeled dataset is partitioned into a training set, a validation set, and a test set. The model parameters are learned during training and the model is further hyperparameter-tuned during the validation stage. Proper tuning ensures optimal fit of the data, improving the likelihood that the model will generalize to new input data. (D) Model performance is evaluated using metrics, such as accuracy, precision, and recall, which are calculated from the number of true and false predictions organized by a confusion matrix. Receiver operating curves (ROCs), which plot the true positive rate against the false positive rate, and precision-recall (PR) curves, which plot precision against the recall, are commonly used to evaluate model performance. A perfectly performing model has an AUC-ROC of 1.0. Illustrated is a progression of AUC-ROC with improving performance (gray being the worst performance and green being the best performance).

be applied if sufficient quantities of data are available. For binarization tasks, the threshold is defined by the researcher to appropriately select for a desired property. For instance, bacterial growth inhibition can be measured by optical density, with compounds resulting in >80% growth inhibition being defined as growth inhibitory (labeled 1) and compounds resulting in <80% growth inhibition defined as not growth inhibitory (labeled 0). The model can then predict a novel molecule not seen during training as potentially bioactive (prediction values approaching 1) or potentially not (prediction values approaching 0). It is up to the researcher to determine how to interpret the model prediction scores based on the specific task.

As a concrete example, Liu et al.[35] developed a message-passing deep neural network (MPNN) to predict the growth inhibition properties of molecules against *Mycobacterium tuberculosis*. An MPNN automatically transforms the graph structure of a molecule into a continuous vector. The authors trained their model on the inhibitory activity of ∼50,000 chemicals against 152 *M. tuberculosis* mutant strains. Growth inhibition was represented as a Z-score of the natural log fold change of the mutants relative to their growth in a control solvent and binarized using a calculated threshold of −4, such that compounds with inhibitory activity (a Z-score less than or equal to −4) were labeled as 1 and compounds with no activity (a Z-score greater than −4) were labeled as 0. This trained model then successfully predicted the activities of an external set of 44 compounds against their intracellular targets in *M. tuberculosis*.

Where Liu et al. leveraged a message-passing architecture for molecular property prediction, Richter et al.[36] trained a random forest (RF) model to predict compound accumulation in *E. coli*. An RF[37] (Figure 2B) is a multiclass classifier that models every possible event and the associated outcome as branches, then uses learned conditional rules to draw a path down the branches to a final prediction. The authors trained their RF model on the intracellular accumulation data of 68 chemicals, quantified using liquid chromatography with tandem mass spectroscopy. Each compound was represented by a vector containing 297 fixed molecular descriptors that defined physicochemical features, such as rotatable bonds and globularity. Their model predicted that accumulation may be dictated by the presence of primary amines, high rigidity, and low globularity. The authors then used this prediction to develop a broad-spectrum analog of deoxynybomycin, a Gram-positive–specific antibacterial molecule.

Where these two examples dealt with vector representations of small molecules from chemical libraries, we emphasize that a wide array of data types are generated during empirical screening, and there are multiple ways a single type of data can be encoded. For instance, sequences (nucleotide and amino acid sequences) can be transformed into a vector using a one-hot encoding that numerically represents each character of a sequence as its own vector.[38,39] For example, in a genome sequence, the nucleotides can be encoded as A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], and T = [0,0,0,1], where the position of 1 indicates the associated nucleotide.[40] Moreover, images (e.g., bacterial micrographs) can be analyzed as raw pixel intensity, then transformed into a vector after a series of alternating convolutional and pooling steps using a convolutional neural network (CNN; Figure 2B). This process is used to extract the most pertinent features from an input image.[41] Indeed, antibacterial research is diverse, so the "best" data representation is dependent on your specific scientific question.

Model selection

ML algorithms can generally be classified as supervised, unsupervised, or reinforcement learning models (Figure 2B)—we are focusing here on the first two. Supervised learning models use labeled datasets for training and learn associations between each input (chemical structure) and their known output (growth inhibition) using regression (for continuous prediction tasks) or classification (for discrete or categorical prediction tasks). Commonly used classical supervised learning models include RFs, which we briefly described above, and support vector machines (SVMs)[42] (Figure 2B). An SVM is a binary classifier that learns to best separate points of a dataset into two classes using a hyperplane, termed the "decision boundary." A hyperplane is a separating boundary and has one less dimension than that of the data feature space. For example, if the feature space is two-dimensional, the hyperplane will be a one-dimensional line. In contrast to supervised models, unsupervised models are used to identify patterns in unlabeled data. Here, inputs are presented with no labeled output and the model attempts to learn the features of the dataset automatically. These models are often used to cluster data for similarity analyses (Figure 2B), for example, using *k*-means clustering, which partitions a dataset into *k* clusters based on the distance of each datapoint to a center point. Unsupervised learning can also be used for dimensionality reduction (Figure 2B) using principal component analysis,[43] which preserves the most pertinent global features of the dataset by maximizing the variation between the datapoints, or t-distributed stochastic neighbor embedding,[43] which is used to project high-dimensional data onto lower-dimensional space while maintaining local relationships within the dataset.

A subset of ML is deep learning (DL), which applies neural networks to learn (Figure 2B). At a minimum, neural networks have three layers: an input layer, which takes in data that is encoded as a vector; a hidden layer, which extracts features from the input layer; and an output layer, which outputs the prediction(s) defined by the user. At a high level, the connections between the nodes of each layer represent the model's parameters, or "weights," which are optimized during training. Overall, DL tends to be reserved for larger datasets, since it is ideally suited to learn complex relationships among many input and output values. Though a wide array of different neural network architectures exist, some of the most common for the scope of this review include feed-forward neural networks (FFNNs), which are general predictive algorithms[44] that have broad applicability, and CNNs, which work spatially and are typically used with image datasets[45] for various image classification tasks.

Training, validating, and testing a model

Training is the process through which model parameters are learned with the goal of outputting the most accurate predictions for the widest array of inputs. Typically, a given labeled dataset is partitioned into a training set (~80% of the labeled dataset), a validation set (~10% of the labeled dataset), and a test set (~10% of the labeled dataset) (Figure 2C). Separate sets for validation and testing are used to monitor a model's predictive accuracy without directly contributing to the training process. This ensures that the model is not learning parameters specific to a single dataset, thereby improving its ability to generalize to new input data not seen during training.

ML algorithms learn by minimizing a loss function—minimizing the difference between a predicted value for some input and the true value—and optimizing model parameters accordingly. In the context of a DL model, at the onset of training, the neural network will first compute an output value based on the initialized weights (the weights are initialized as random values) in the "forward pass" phase. Once the output is computed, the model will calculate the error (the difference between the predicted output value and the true value), then propagate this error backward through the model to adjust each weight in a process termed "backpropagation." The model will iterate through this algorithm a user-defined number of times until the error—or loss—is minimized. Subsequently, the ML model will be evaluated by performing predictions on the validation set, which provides an opportunity to optimize the model's hyperparameters for further performance gains. Once a model is sufficiently optimized using the training and validation sets, its performance can be quantified using the test set, which the model has not seen, to more accurately understand how the model may perform for real-world tasks.

An accurate ML model must be able to generalize what it has learned during training to novel inputs. Model overfitting occurs when the model is specific to the training data and is unable to generalize to new inputs. This can happen when the model learns irrelevant features—or "noise"—within the training set. Model underfitting occurs when the model fails to learn the defining features of the training dataset (Figure 2C). Both overfitting and underfitting prevent the model from appropriately generalizing to new prediction sets and result in inaccurate predictions. Model averaging is commonly used to build a rigorous model that avoids overfitting and underfitting. The dropout method[46] is one such approach, which involves randomly excluding weights during training. Another method, termed ensembling,[47] compensates for individual model inaccuracies by averaging the outputs of an array of models that have different weights and/or architectures. Ensembling can be performed in a variety of ways, with a common method involving the development of multiple copies of the same architecture, training each on a different subset of the training dataset, and then averaging the predictions thereafter.

Model performance is evaluated using computed metrics, such as accuracy (the ratio of correct outputs to the total outputs), precision (the ratio of correctly predicted positive outputs to the total predicted positive outputs), and recall (the ratio of correctly predicted positive outputs to the total number of real positive observations).[48] These metrics are used for classification tasks and are calculated from the true positives (TPs), false positives, true negatives (TNs), and false negatives, commonly organized using a confusion matrix (Figure 2D). Their use and interpretation is heavily dependent on the goals of the prediction task, and the uncertainty costs must be considered accordingly.[49] For example, prioritizing precision might be beneficial when screening expensive compounds since false-positive predictions may be prohibitively costly. Other standard performance measurements include the F1-score, which aims to address tradeoff issues inherent in precision and recall by considering both together; receiver-operating characteristic (ROC) curves (Figure 2D), which plot the true-positive rate against the false-positive rate at different confidence cutoffs; and precision-recall (PR) curves, which plot the precision and recall at different confidence cutoffs. The areas under these curves (AUC) are used to quantify model performance. ROC curves are typically used with balanced datasets where we observe equal numbers of positive and negative outputs, whereas PR curves are used when there is high-class imbalance—vastly different numbers of positive and negative examples from a given training dataset. Although many methods exist to train ML models, a common sentiment is that 80% of the acquired dataset should be centered around collecting and preprocessing data for training and the remaining 20% should be reserved for model optimization and testing.[50] As mentioned, given the complexity of chemical and biological systems, collecting large and high-quality datasets is essential to minimize prediction error and maximize generalization in real-world settings.[51]

REVISITING THE WAKSMAN PLATFORM WITH ML

The majority of our clinically used antibiotic classes are naturally occurring molecules discovered during the golden era of antibiotic discovery, which lasted from the 1940s to the mid-1960s[52] (Figure 1). The great success of this period was due to the invention of one of the early systematic screening methods for antibiotics, the Waksman Platform. This approach involved isolating secondary metabolites from soil-dwelling microbes, the most common being the Actinomycetes, and screening these for antibacterial activity *in vitro*. NPs are structurally and functionally diverse[53] and their physicochemical properties have been honed for antibacterial activity over billions of years of evolution, making them potent and highly precise antibacterial compounds.[54] Although the Waksman platform dominated for nearly two decades, it suffered from laborious isolation and purification steps, as well as a persistent rediscovery problem, wherein similar compounds were repeatedly being isolated.

After somewhat of a hiatus between the 1970s and early 2000s, the renaissance of NP discovery efforts in contemporary antibacterial research is a result of progress in our understanding of the metabolic pathways encoded by biosynthetic gene clusters (BGCs) that govern NP production.[55] Indeed, technological advancements, such as next-generation sequencing, have allowed for the collection and annotation

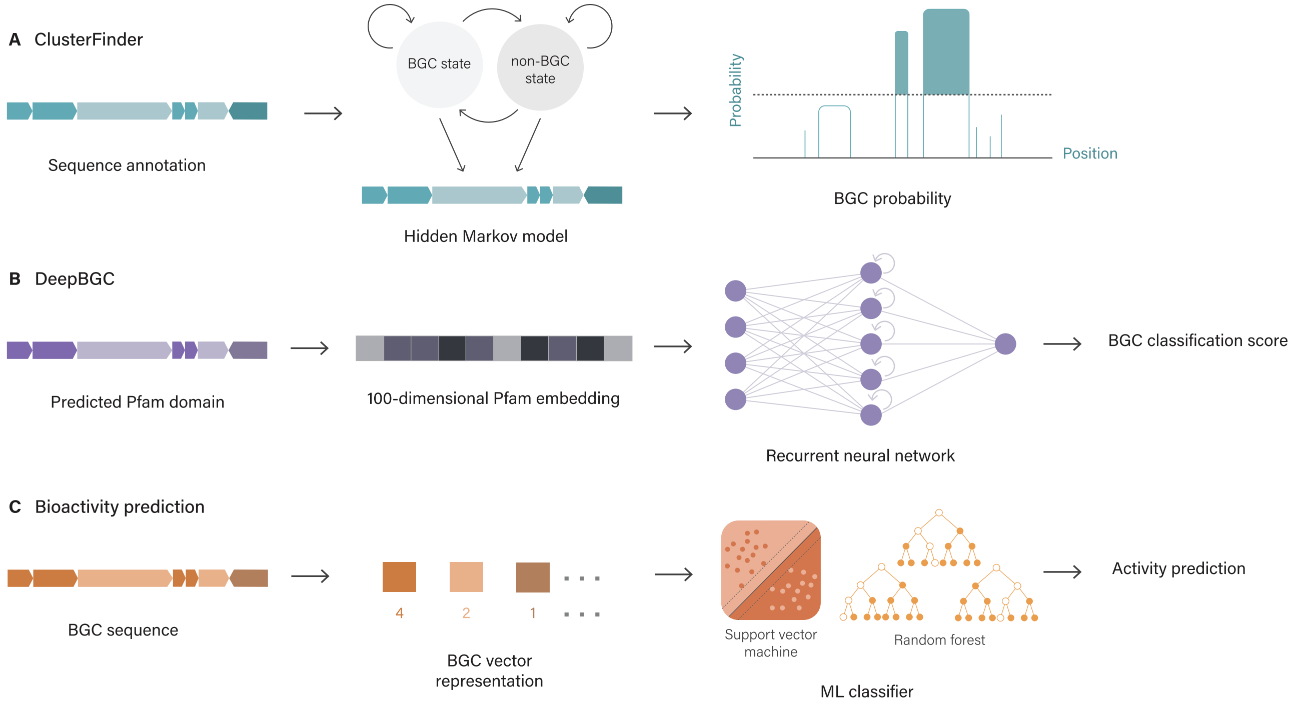

FIGURE 3 Predicting natural products using machine learning. (A) ClusterFinder uses an HMM to identify new BGCs from metagenomics datasets. HMMs are probabilistic models that can classify sequences based on their underlying hidden structure (see text). In this study, an HMM predicts the probabilities that a Pfam domain belongs to a BGC and a non-BGC. Sequences with domains that have a high predicted probability of belonging to a BGC are clustered and annotated. (B) DeepBGC takes as input NLP-based word embeddings of Pfam domains that can preserve their structural and sequential relationships. RNNs are neural networks that remember their input using an internal "memory" state, making them ideal for capturing the order of sequence data. The model outputs a BGC classification score for an inputted Pfam domain. (C) ML-based models have been developed for bioactivity prediction solely based on the sequence. A classifier developed by Walker and Clardy takes as input a BGC sequence, transforms it into a vector representation based on the number of different domains occurring within the sequence, and then outputs a bioactivity prediction.

of whole genome sequences of common NP producers, as well as large metagenomics databases of uncommon producers.[56,57] Together, these advancements catalyzed the development of computational genome exploration tools for identifying novel BGCs. Contemporary NP discovery pipelines now consist of four general stages: (1) genome annotation; (2) BGC identification; (3) dereplication and structure elucidation; and (4) antibacterial activity profiling.

Genome mining tools for BGC detection, such as CLUSEAN[58] and antiSMASH,[59] began to emerge in the early 2000s. These were homology-based methods that used expert-defined rules and reference alignments to identify BGCs.[60,61] These approaches were important for initiating the resurgence of NP development in antibiotic research after the golden era, and they continue to be widely used. However, these homology-based models are not ideal to detect the chemical diversity inherent to NPs. For example, a challenge frequently encountered with nonribosomal peptide discovery is the inability to detect modifications, such as glycosylation, which occur after assembly by the nonribosomal peptide synthetase.[62] Homology-based models are also limited in their ability to discover structurally novel secondary metabolites.[63] Fortunately, ML models are well-positioned for generalizing across large BGC sequence spaces.[64] As a concrete example, ClusterFinder[65] is a hidden Markov model (HMM) developed for novel BGC identification from metagenomics datasets (Figure 3A). HMMs are probabilistic models that can implicitly analyze underlying hidden patterns in datasets based on some given observable condition. For example, HMMs can be used to predict the location of substructures and motifs within a sequence by using probabilistic models of substructures or features, whose parameters are learned from the composition of the sequence (e.g., frequencies of various DNA sequence patterns in the case of DNA sequences as input). ClusterFinder uses an HMM to predict the probabilities that a protein family (Pfam) domain belongs to a BGC and a non-BGC. These probabilities are based on the domain's frequency in a BGC, a non-BGC, as well as the frequency of its neighboring domains. Sequences with domains that have a high predicted probability of belonging to a BGC are then clustered and class annotated. For this purpose, ClusterFinder was trained on 677 experimentally characterized BGCs and 100 non-BGCs to learn the probability that a Pfam domain belongs to a BGC or a non-BGC. These probabilities were calculated using the frequency of occurrence of the domains within the training set sequences. ClusterFinder was then used to predict with high confidence ~11,000 BGCs in ~1000 different genomes, of which 69% were not identified by rule-based methods, such as antiSMASH.[65] More recently, ClusterFinder was used to discover the thiopeptide antibiotic lactocillin from the human microbiome.[66]

Genome mining was further improved through the adoption of natural language processing (NLP)-based DL models that equate genetic sequences to language—both genetic sequences and language are based on strings of characters that follow their own syntax in order to be functional.[67] Therefore, NLP models can preserve the underlying relationships of a sequence (e.g., gene sequence structure) in a vector representation, a challenge that limited BGC identification with other models, particularly in the context of generalization to new sequence space. For instance, Hannigan et al.[68] adopted an NLP-based approach to transform Pfam domains into word embeddings to develop DeepBGC, a recurrent neural network (RNN) designed to identify novel BGCs (Figure 3B). RNNs are neural networks that can remember their input using an internal "memory" state, making them ideal for capturing the order of sequence data. DeepBGC was trained on a diverse dataset of ∼600 known BGCs and ∼10,000 non-BGC sequences. The model predicted a BGC classification score for inputted Pfam domains encoded using word embeddings. DeepBGC was evaluated against, and outperformed, ClusterFinder across three tasks: (1) locating BGCs within whole genomes (DeepBGC: AUC = 0.923 and ClusterFinder: AUC = 0.847); (2) differentiating BGCs and generated non-BGCs (DeepBGC: AUC = 0.984 and ClusterFinder: AUC = 0.936); and (3) locating novel BGC classes from a test set (DeepBGC: AUC = 0.946 and ClusterFinder: AUC = 0.865).

Integration of nongenomics experimental data with ML has been successful in the later stages of NP discovery, including dereplication, an essential and laborious procedure used to filter known secondary metabolites. Dias et al.[69] developed an NP discovery approach to predict compounds with antibacterial activity, with the objective of reducing time and biological activity screening costs associated with conventional NP discovery pipelines. Here, the authors trained a variety of ML and DL models (e.g., RFs, SVMs, and CNNs) to learn NP structure–activity relationships from a collection of 116 crude extracts, fractionated extracts, and purified compounds. The molecules were encoded using descriptors derived from źH and ^{13}C NMR spectra. The authors validated their model using an external test set of four new compounds and empirically tested their results *in vitro*. This led to the discovery of a novel NP that was able to inhibit the growth of methicillin-resistant *Staphylococcus aureus*.

Additionally, ML approaches have been used for bioactivity prediction. Walker and Clardy[70] developed an ML-based approach to predict NP antifungal and antibacterial activity from BGC sequence alone (Figure 3C). They trained three ML classifiers (SVM, RF, and logistic regression) with a curated set of known BGC sequences, coupled with the binarized bioactivities (antibacterial, antifungal, anticancer, and cytotoxic) of their encoded NP. The objective of their approach was to predict whether an inputted BGC sequence encoded a bioactive NP. To evaluate their model's performance, they computed balanced accuracy (the average of the TP and TN rates), ROC curves, and PR curves for each ML classifier. The authors assessed their models' performances using a test set of 258 BGC sequences that varied in their similarity to the training set. The best classifier had an accuracy of ∼80% and the worst was ∼60%. Of note, BGCs that were similar to those in the training set were classified with greater accuracy, emphasizing the need for larger and more diverse datasets during training to improve generalization across wider regions of sequence space.

Many BGCs encode for secondary metabolites that are not synthesized in standard laboratory conditions, but may be induced through culturing in unconventional growth media, as well as through heterologous expression using synthetic biology approaches.[71] To this end, modern NP discovery makes use of biological engineering principles to regulate and optimize NP biosynthetic pathways. Indeed, ML and DL approaches have shown promise in assisting with promotor prediction,[72] promotor design,[73,74] codon optimization,[75] and protein engineering.[76] For example, Kotopka and Smolke[73] exemplified how DL can be used to predict promoter activity and generate novel ones. The authors trained a CNN on promoter sequence–activity relationships collected from the gene expression activity of 675,000 sequences in the yeast *Saccharomyces cerevisiae*. Though they are commonly used with image datasets, CNNs are ideal for use with sequences since they can preserve their spatial structure. The model took as input matrix representations of gene sequences (transformed using one-hot encoding) and predicted their promoter activity. Using this large training set, they created new promoter sequences using three generative models, with the objective of maximizing promoter activity, and synthesized them for empirical validation. First, in a screening design strategy, randomly generated promoter sequences were taken as input for their CNN. If the predicted activity surpassed a specified threshold, the promoter was shortlisted for synthesis. Second, using an evolution strategy, the authors generated promoter sequences that were iteratively mutated and evaluated for desired activity *in silico* until the activity surpassed the threshold value. Lastly, using a gradient ascent design strategy, randomly generated promoter sequences were altered until their predicted activity score was optimized. The authors successfully designed and experimentally validated over 100 promoters and, using this multipronged approach, found that promoters designed using the evolution or gradient ascent strategies generally exhibited a greater or similar level of activity to their benchmark dataset. Overall, NP discovery will greatly benefit from the curation of large, standardized datasets. This will accelerate the development of new ML models, improve their ability to generalize to new sequences, and broaden the utility of synthetic biology methods toward NP production.

EXPANDING THE CHEMICAL SEARCH SPACE FOR NEW ANTIBIOTICS

As the golden era waned toward the mid-1960s, medicinal chemistry efforts grew in response (Figure 1). In an effort to bypass resistance determinants, improve absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties, and increase antibacterial potency, existing NPs were used as scaffolds toward the development of semisynthetic derivatives thereof.[77] This era was highly successful in these efforts. For example, medicinal chemistry approaches were responsible for the optimization of nalidixic acid into fluoroquinolones. However, it was also marked by a general lack of innovation, owing to the challenge of developing truly novel classes of antibiotics.[4]

Occurring somewhat in parallel with the medicinal chemistry era of antibiotic discovery was the emergence of high-throughput screening of synthetic chemical libraries. Here, the development of combinatorial chemistry methods and improved laboratory automation enabled the assembly and manipulation of large chemical collections that frequently exceeded hundreds of thousands of compounds. These technological advancements were complemented by the development of genomics technologies, such as recombinant DNA methods, to identify and purify biologically "essential" protein targets for chemical screening.[17] The convergence of innovation in chemistry, laboratory automation, and genomics enabled the growth of target-based screening as a dominant antibiotic discovery model. Though this approach resulted in many hit molecules, the lack of diversity of chemical libraries used, as well as the reductionist nature inherent to *in vitro* assays, limited the discovery of viable whole-cell–active antibiotics.[4] We note that synthetic compound libraries are typically composed of chemicals with drug-like properties.[78] However, antibiotics have unique physicochemical properties relative to conventional human drugs (greater molecular weight, chemical complexity, polarity, and charge) and, therefore, exist in a different chemical space.[79] As such, typical synthetic compounds are generally less suited to evade efflux or bypass permeability barriers, such as the Gram-negative outer membrane.[80]

More contemporary high-throughput screening programs for antibacterial molecules are conducted using live cells,[81] which partially addresses the aforementioned issues encountered in target-based *in vitro* screens. Unfortunately, attrition rates in the antibiotic development pipeline have not sufficiently waned,[7] and developing an antibiotic remains expensive and challenging.[82] Indeed, in addition to antibacterial efficacy,[79,83] an antibiotic must have additional pharmacological properties that permit clinical efficacy.[78] As a result, a promising antibacterial compound may spend years in early development stages, before ultimately being deemed unsuitable due to concerns of toxicity, PK/PD, and/or other ADMET properties.

At a high level, antibiotic drug discovery is a multiproperty optimization problem. A primary goal of preclinical antibiotic discovery should be in predicting the success of a compound as early in the development process as possible. Unsurprisingly, this is no trivial feat; it has been estimated that 10^{60} drug-like chemicals can theoretically exist.[84] Fortunately, such multiproperty optimization tasks are ideally suited for ML and DL methods since they can rapidly explore broad chemical search spaces and enrich for the most promising compounds for downstream validation and optimization.[85]

Quantitative structure–activity relationships (QSARs) are frequently used in drug design and for assembling chemical libraries.[86,87] Contemporary QSAR modeling is largely derived from the work of Hansch et al.,[88] and, in the early 1980s, QSAR models began to focus on bioactivity prediction in drug discovery.[89] ML algorithms have since become a staple in QSAR modeling, where representations of chemical structures are built from manually curated parameters of a molecule's structure or physicochemical properties, such as hydrophobicity, molecular weight, and the number of rotatable bonds (Figure 4A). For example, Ivanenkov et al.[90] trained an array of ML models, including *k*NN, RF, SVM, and FFNN (Figure 2B), on a standardized dataset of 73,000 molecules to predict compounds with activity against *E. coli*. For reference, this highly imbalanced training dataset contained 8724 active compounds and 65,843 inactive compounds. The authors used 40 molecular descriptors for their compounds, including hydrophobicity, number of hydrogen bond donors and acceptors, number of rotatable bonds, and others. They assessed their model using an external test set of 5000 compounds that were randomly selected for low similarity to the training set. This model predicted 371 active compounds, of which 13 exhibited considerable activity when validated *in vitro* against *E. coli*.

Although classic molecular fingerprint–based ML models can perform adequately when performing predictions in well-characterized chemical spaces, they are unable to generalize well to novel chemical spaces due to the inherent limitation of using human-defined features. QSAR models greatly benefited from computational advancements that enabled the development and widespread use of graph neural networks (GNNs)[91,92] (Figure 4A). Chemprop[93] is considered a world-leading GNN-based molecular property predictor. Chemprop applies an MPNN architecture to aggregate information about local molecular structure into a single vector representation, which is then taken as input into an FFNN for molecular property prediction. For instance, Stokes et al.[18] trained Chemprop on a collection of ∼2500 small molecules for those that were growth inhibitory against *E. coli*. Growth inhibitory activity was binarized as 1 (inhibitory) or 0 (not inhibitory). Following training, their model was used to virtually screen a collection of over 100 million structurally diverse compounds[94] for antibacterial activity against *E. coli*. From this screen, the authors discovered halicin, a novel antibacterial molecule with activity against an array of diverse pathogens, including *Acinetobacter baumannii*, *Clostridium difficile*, and *M. tuberculosis*.

DL approaches can also be used to traverse uncharted chemical space through *de novo* antibiotic design with directed molecular generation using so-called generative ML models. For example, variational autoencoders (VAEs)[95,96] use two neural networks (together called an autoencoder): one to encode molecular structures into dense vectors and one to decode the vectors back to a molecular structure (Figure 4B). This compression and decompression of information allows the model to learn molecular features. New molecules can be generated by VAEs through the addition of noise to their latent vector representation. Gómez-Bombarelli et al.[97] developed a VAE for the molecular generation to design molecules optimized for their synthetic accessibility score (SAS) and their drug-likeness, as defined by their qualitative estimate of drug-likeness (QED)[98] score. They trained their model on SMILES strings of 250,000 drug-like molecules from the ZINC database.[94] The VAE was then combined with a molecular property predictor to evaluate VAE output molecules *in silico*. Their model was able to take as input SMILES strings, encode them into a latent representation, and iteratively optimize the compounds for the desired properties. Though their model was consistent in its ability to

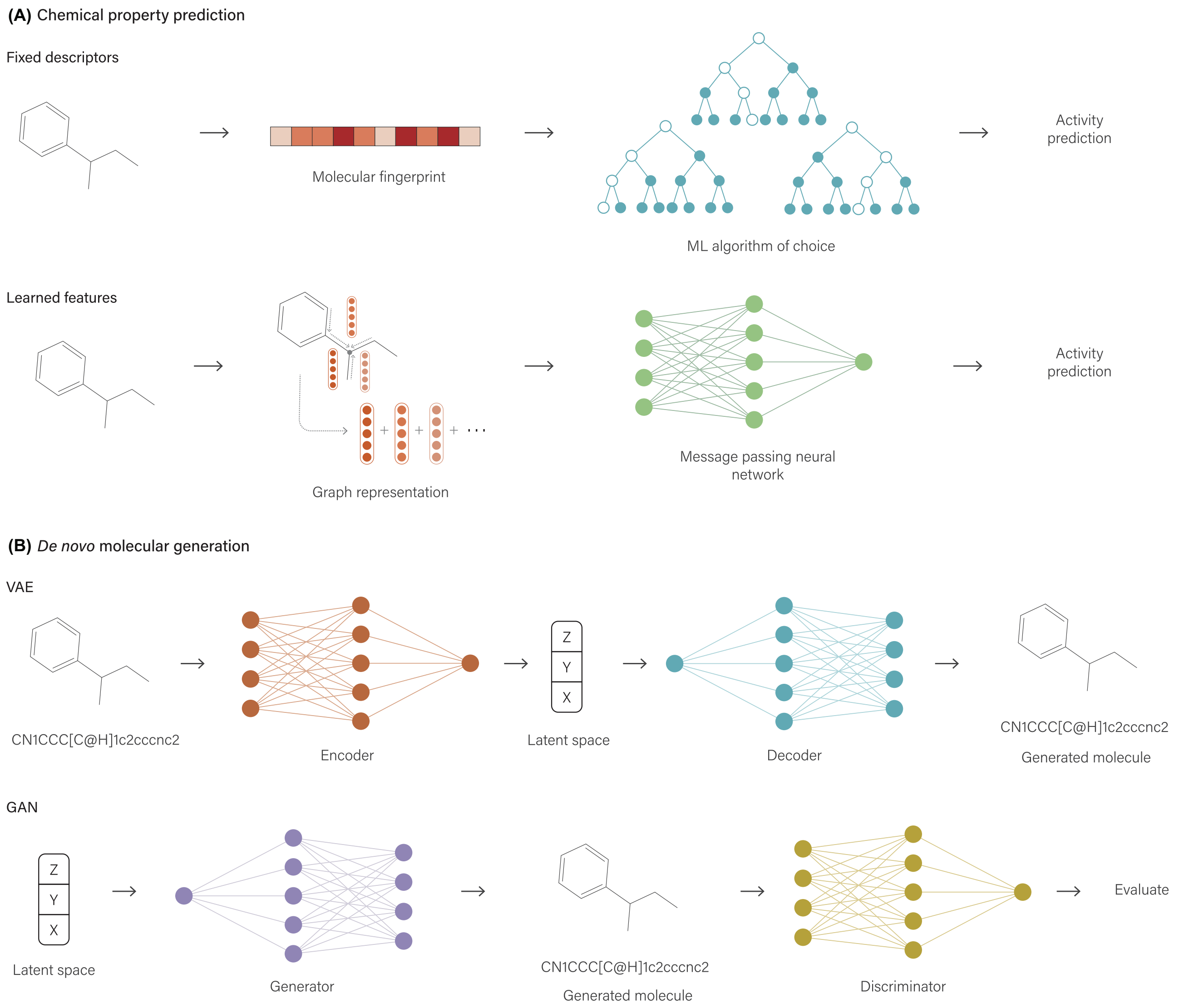

FIGURE 4 Machine learning for small molecule antibiotic discovery. (A) ML algorithms are useful for exploring antibacterial chemical spaces. ML algorithms can be trained using classical molecular fingerprints, which are expert-defined parameters of a molecule's structure and physicochemical properties, such as hydrophobicity, molecular weight, number of rotatable bonds, among others. GNNs automatically generate vector representations of molecular graphs and can more accurately generalize to novel chemical spaces. (B) DL algorithms can also explore vast regions of chemical space through *de novo* molecular generation. VAEs and GANs are two popular architectures that have enabled progress in *de novo* molecular generation for various molecular design tasks.

generate realistic molecules with improved SAS and QED scores from the lowest-scoring molecules from the ZINC database (molecules in the 10th percentile), the authors found that many of their generated molecules violated stability and synthetic constraints, rendering them unviable.

Jin et al.[99] expanded on the conventional VAE with the development of the junction tree variational autoencoder (JT-VAE) which creates a vocabulary of extracted molecular substructures that can be used to generate chemically valid novel compounds, thus overcoming the problems encountered by Gómez-Bombarelli et al. With a JT-VAE, a molecule is generated through the iterative addition and scoring of individual nodes that correspond to molecular substructures. The authors trained their model on a set of ~250,000 molecules from the ZINC database,[94] then evaluated the ability of their model to generate molecules with a desired octanol-water partition coefficient (logP) value and optimize a given molecule to improve its logP value. The JT-VAE outperformed other variations of the VAE in its ability to generate and optimize molecules for a desired logP; each of its top three generated molecules obtained higher property scores, and their model was able to improve a given molecule's logP while still maintaining the new molecule's similarity to the original compound, with a success rate of 83.6%. Importantly, all JT-VAE-generated molecules were chemically valid, thus outperforming prior generative models.

Another approach to *de novo* molecular generation applies generative adversarial networks (GANs)[100–103] (Figure 4B). At a high level, these models place two neural networks in competition with one another: a generative network creates molecules, while an adversarial network evaluates if those molecules are real (from the training set) or generated. The two networks learn from each another until the generative network can create novel compounds that are indistinguishable from the training set. Méndez-Lucio et al.[103] were able to generate molecules by training two conditional GANs on a publicly available dataset of over ∼20,000 compounds and their corresponding gene expression profiles (978 genes) from chemically perturbed cancer cells.[104] Their objective was to design a molecule that could produce a desired gene expression profile. Here, a generative network takes in as input a desired gene expression profile and a random noise vector to create a novel molecule. Next, an adversarial network evaluates if the generated molecule is "real" (from the training dataset) or generated. A conditional neural network then determines the probability of the generated molecule eliciting the required transcriptional response. This conditional network takes as input both the gene expression fingerprint and the vector representation of the generated molecule and assigns a classification score to the molecule (the higher the score, the more favorable the molecule). This process is repeated with a second conditional GAN to refine the properties of the novel molecule. Although their model was successful in its objective, it was not highly generalizable and failed to generate structurally novel compounds relative to what was observed during training. Nevertheless, their study begins to highlight the utility of ML approaches in designing compounds with specific targets[17,103,105,106] or with multiple targets[107–110] in a specific biological system. Approaches such as this may be adopted in antibiotic discovery for similar purposes as in cancer research.

Antibiotic discovery is a multiproperty optimization problem, so it is ideally suited for ML techniques. Multitask prediction models[111,112] identify features associated with multiple properties to enable the prediction of molecules with the pharmacological features necessary for human use, while maintaining antibacterial efficacy. Though multitask prediction models remain an emerging trend in drug discovery, some studies have shown their early potential.[113–115] For example, Khemchandani et al.[114] developed DeepGraphMolGen, a multitask model to generate molecules that are able to bind to dopamine transporters, but not norepinephrine transporters. Their approach consists of two phases: property prediction and molecular generation. In the first phase, a GNN and an FFNN are used to learn the features inherent to molecules with a high binding affinity to dopamine transporters and a low affinity for norepinephrine transporters. To this end, they trained their model on publicly available datasets of molecular binding affinities for each transporter (4506 values for dopamine and 2780 values for norepinephrine).[116] After training their model, a reinforcement deep learning model (RL) was used to generate molecules in the second phase. An RL model generates a molecular graph by iteratively adding a bond or substructure, beginning with a single atom or molecule. We note that an RL model learns using rewards to guide the design process: it is rewarded for generating a SMILES string that is valid, drug-like, can be synthesized, and which optimizes the desired computed binding constants. The generated molecule is then inputted to the GNN to predict the generated molecule's binding constant to both dopamine and norepinephrine transporters. The cycle continues until the molecule is optimized. Although their findings were not experimentally validated, the authors present a promising approach to the multitask molecular generation that may be used toward *in silico* antibiotic drug development.

EMERGING APPLICATIONS OF ML IN ANTIBIOTIC DISCOVERY

Systems biology

Modern antibiotic drug discovery consists of a diverse suite of approaches, including unconventional methods, such as target-enriched whole-cell screening, testing naïve molecules in physiologically relevant conditions, and antibiotic adjuvant development.[117] This methodological diversity reflects the newfound appreciation for the intrinsic complexity of antibiotic function in the context of the whole cell, which has resulted in the adoption of systems-level approaches that integrate chemical biology with genome-wide investigations.[15,16,118,119] However, these methods also tend to generate dense, multidimensional datasets that are challenging to comprehensively interpret due to the complexity inherent to biological systems.

Fortunately, ML approaches may be used to deconvolute causal relationships between chemical or genetic perturbations and phenotypic outcomes, as well as help us elucidate cryptic aspects of bacterial physiology that can be leveraged toward new antibiotic development.[120–122] For example, Kavvas et al.[120] developed Metabolic Allele Classifier (MAC), a flux balance analysis–based ML classifier, to predict resistance mechanisms arising in *M. tuberculosis* against seven diverse antibiotics—isoniazid, rifampicin, d-cycloserine, ofloxacin, ethambutol, pyrazinamide, and para-aminosalicylic acid. The model is trained on: (1) genome sequences of *M. tuberculosis*; (2) their corresponding resistance phenotypes; and (3) a genome-scale metabolic model of *M. tuberculosis*. The authors used a training set consisting of 375 susceptible and resistant isolates of *M. tuberculosis*. The model takes as input the gene sequence of an *M. tuberculosis* strain and provides an interpretable prediction of its resistance phenotype in the form of the associated genetic and metabolic determinants (given as steady-state fluxes). The authors showed that, in addition to a high classification accuracy (AUC = 93% for isoniazid), MAC was able to identify both known and novel genetic resistance determinants. It could also identify metabolic pathways that differed between resistant and susceptible strains for para-aminosalicylic acid, pyrazinamide, and isoniazid.

Additionally, Woo et al.[123] developed Deep gene COmpound Profiler (DeepCOP), an FFNN to predict the up- and downregulation of genes of cancer cells following a chemical perturbation (Figure 5). The authors trained DeepCOP on chemical–genetic fingerprints derived from a collection of 1.3 million gene expression profiles of 77 different cancer cell lines chemically perturbed with ∼20,000 compounds.[104] Gene expression was represented as standardized Z-scores and

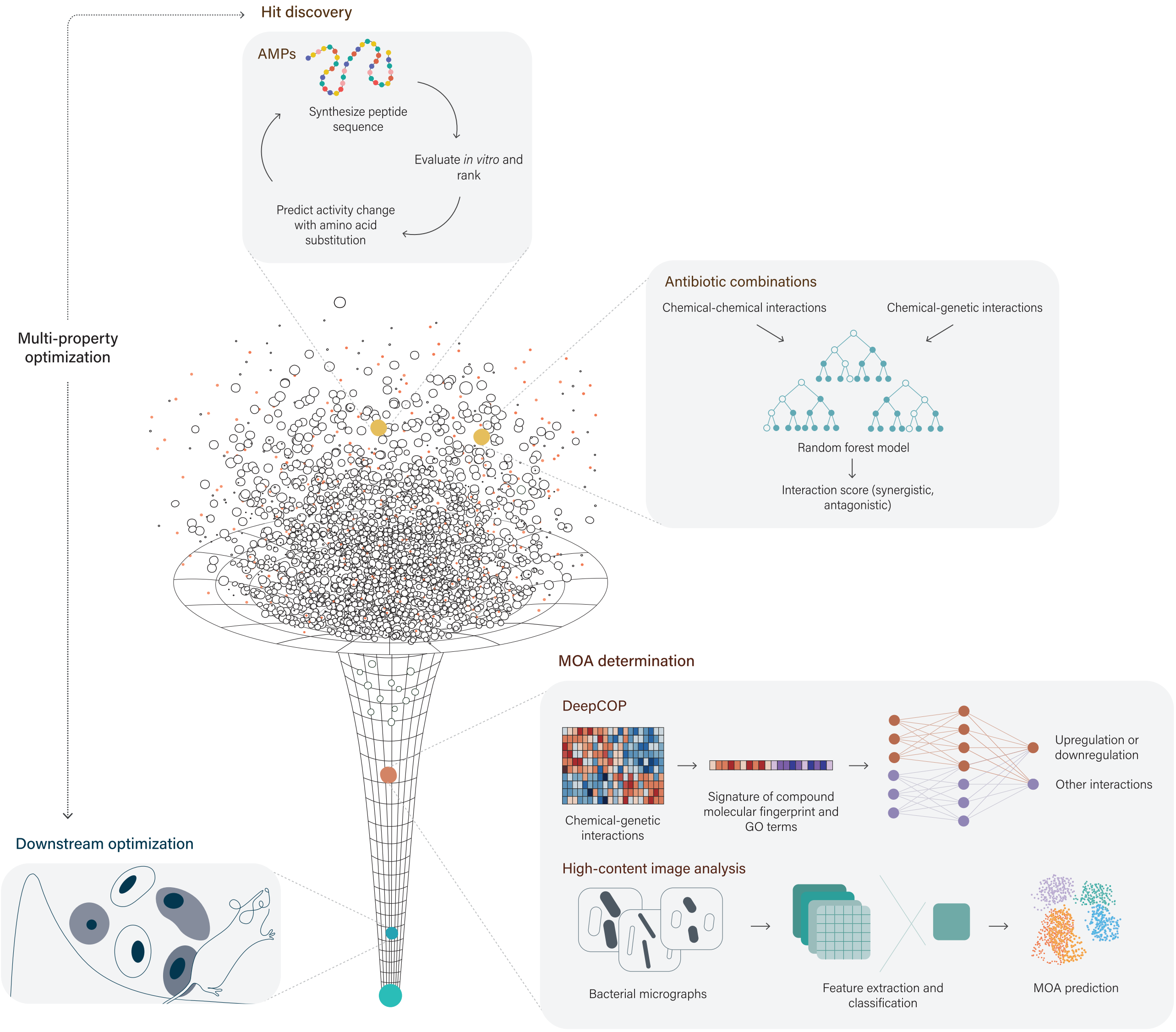
Hit discovery
AMPs
Synthesize peptide sequence
Evaluate in vitro and rank
Predict activity change with amino acid substitution
Multi-property optimization
Antibiotic combinations
Chemical-chemical interactions
Chemical-genetic interactions
Random forest model
Interaction score (synergistic, antagonistic)
MOA determination
DeepCOP
Chemical-genetic interactions
Signature of compound molecular fingerprint and GO terms
Upregulation or downregulation
Other interactions
High-content image analysis
Bacterial micrographs
Feature extraction and classification
MOA prediction
Downstream optimization

FIGURE 5 Emerging challenges for machine learning in antibiotic discovery. The major stages in preclinical antibiotic discovery, highlighting how emerging ML and DL approaches can be integrated end-to-end to streamline new antibiotic discovery and development. ML-based approaches allow the exploration of vast chemical and sequence space, and, in combination with multiproperty optimization methods, can reduce attrition rates in the later stages of antibiotic development by increasing the probability of identifying promising candidates.

binarized using upper and lower threshold values; upregulated genes were above the upper threshold and downregulated genes were below the lower threshold. Chemical structures were represented using rule-based molecular fingerprints, and each gene's corresponding gene ontology (GO) terms were vectorized using one-hot encoding. The authors concatenated the molecular fingerprints with the GO term vector representation and used these combined vectors to train their FFNN. Since genes within the same pathway or with the same functional class share GO terms, this allowed the model to learn chemical–gene interactions which it could use to make predictions on the regulation of genes in response to an unknown compound. They validated their model (AUC ~0.8) by comparing its predictions of endpoint gene expression for three compounds outside of the training dataset (enzalutamide, VPC-17005, and VPC-14449) to the experimentally derived RNA-Seq profiles. While the focus of this early proof-of-concept study was cancer biology, similar approaches can be adopted for developing antibacterial compounds that induce a desired gene expression profile, as well as in our efforts to develop narrow-spectrum antibiotics by exploiting genetic or metabolic pathway differences at the species[124–126] or genus level.[127,128]

Combinatorial therapies

Chemical–genetic interactions form the biological basis of combinatorial therapies, wherein multiple compounds are administered together with the goal of observing synergy—the instance where a combination of molecules shows greater bioactivity than the expected sum of their individual activities.[129] As highlighted above, ML may be integrated with systems-level approaches to elucidate dense chemical–genetic networks underlying drug synergy phenotypes. ML may also aid in the discovery of novel antibiotic combinations since systematically exploring antibiotic combinations at a large scale is not feasible, owing to the combinatorial explosion that arises when empirically testing pairwise chemical combinations. For example, with 10 compounds, the number of possible pairwise combinations equals 45. With 1000 compounds, a relatively small chemical collection, the number of possible pairwise combinations equals 499,500. Therefore, ML approaches are ideally suited for predicting synergistic antibiotic combinations, given their ability to rapidly traverse through large chemical and biological spaces when appropriately trained.

ML methods for predicting synergistic chemical combinations typically combine experimentally determined drug–drug interactions (fractional inhibitory concentration,[130] Bliss scores,[131] and Loewe additivity[131]) with the molecules' underlying biochemical target or pathway interactions.[132] Chandrasekaran et al.[133] developed INDIGO, an RF model trained on binarized chemical–chemical interaction data—labeled as either synergy or antagonism—as well as previously published chemical–genetic interaction data for *E. coli* gene deletion mutants in response to 15 drugs[134] (Figure 5). INDIGO was trained on chemical–chemical interactions and chemical–genetic interactions, so it could learn the genes that are predictive of synergy or antagonism. Chemical–chemical interactions were represented with their Loewe additivity score and the chemogenomic profiles were binarized based on each strain's sensitivity to a certain compound. INDIGO can then predict synergistic or antagonistic interactions between two compounds using their individual known chemical–genetic profiles. The authors experimentally validated two predicted combinations (fusidic acid with rifampicin and fusidic acid with vancomycin). Mason et al.[135] took an alternate approach to chemical genomics–based modeling and developed CoSynE, an RF model trained on a total of 153 chemical–chemical Loewe additivity interaction scores and rule-based structural fingerprints generated for each compound. Unlike INDIGO, CoSynE does not use any fitness data—only the interaction scores and the corresponding chemical structures. The authors assessed their model using a validation set of six compounds (kanamycin, penicillin, roxithromycin, 5-fluorouracil, mupirocin, and pentamidine). CoSynE predicted 12 synergistic combinations and 10 of these were successfully validated *in vitro*. Other papers that applied similar model architectures for chemical–chemical interaction prediction are referenced here.[136–138]

Unfortunately, the use of DL models in antibiotic combination discovery is not yet common, largely due to a lack of available chemical–chemical interaction data in bacteria at a sufficient scale. Other fields, particularly anticancer drug discovery, have publicly available datasets[139] that have resulted in a comparatively large effort to develop DL models for predicting synergistic drug combinations.[140] For example, Preuer et al.[141] developed DeepSynergy, an FFNN used to predict anticancer drug synergy. Here, the authors trained their model on over 20,000 chemical–chemical interactions (583 pairwise combinations at four different concentrations) in 39 cancer cell lines.[142] Compounds were encoded using rule-based fingerprints of over 4000 descriptors (molecular weight, number of rotatable bonds, electrotopological states, and topological descriptors). Unlike many other synergy prediction models, DeepSynergy does not use chemical–genetic interaction data, instead relying on a published transcriptomic dataset of gene expression profiles for untreated cells from each cell line,[143] thereby allowing their model to differentiate between cell lines. As input, the model receives the gene expression profiles of an untreated cell line and the chemical structures of each drug in the combination. It outputs the predicted synergy score for the two molecules based on the Loewe additivity model. DeepSynergy outperformed ML-based approaches that used RFs and SVMs, achieving an AUC of ∼0.90. Of note, similar DL models have been developed that also use large chemogenomic datasets.[141,144,145]

Jin et al.[146] developed ComboNet, a GNN to predict synergistic drug combinations against SARS-CoV-2. ComboNet is trained using publicly available datasets measuring the reversal of SARS-CoV-2–induced cytopathic effect in Vero E6 host cells. Their training set included: (1) 88 drug–drug interactions; (2) the antiviral activities of over 8000 compounds; and (3) the molecular structures of each compound in the training set. Due to an insufficient amount of drug–drug combination data for SARS-CoV-2, the authors also trained their model on the drug–target interactions of over 10,000 compounds against the SARS-CoV-2 protein targets ACE2, 3CLpro, and Spike-ACE2. To further expand their training set, the authors included data on HIV since the host–viral interactions between HIV and SARS-CoV-2 are sufficiently similar. This included: (1) 114 drug–drug interactions; (2) the antiviral activities of 30,000 compounds; and (3) drug–target interaction data for six HIV targets (HIV-1 protease, integrase, reverse transcriptase, CCR5, CXCR4, and CD4). ComboNet was validated using a prediction set of 153 diverse compounds, which resulted in 11,600 predicted combinations. The 30 highest-scoring combinations were then tested *in vitro* using a SARS-CoV-2 infection assay in Vero E6 cells, and the synergistic interactions of two combinations (remdesivir-reserpine and remdesivir-IQ1S) were observed. Further analysis on the benefit of multidisease training with ComboNet showed that removing the additional HIV data from the training set decreased the ROC-AUC from 0.820 to 0.658. With this study, Jin et al. exemplify a unique method of how a lack of drug–drug interaction data for model training may be overcome in some instances—particularly when training data exist for a sufficiently similar biological system. Overall, the further development of sufficiently large chemical–chemical datasets for an antibiotic activity will help accelerate the development and real-world use of synergy prediction models.

Antimicrobial peptides as emerging antibiotics

Non–small-molecule antibiotics—antimicrobial peptides (AMPs) for example—are becoming recognized as alternatives to conventional antibiotics. Similar to combination therapy development, AMP discovery is an ideal challenge for ML approaches since the identification and classification of AMPs from large sequence spaces is one of the major bottlenecks in their development.[147] A variety of ML model architectures have been used to predict and design AMPs, typically training models on both sequence information and antimicrobial activity data. For instance, Yoshida et al.[148] trained an ML classifier with an evolutionary algorithm on empirically collected AMP IC_{50} values against *E. coli* to optimize their antibacterial activity. Evolutionary algorithms optimize peptide sequences through iterative mutation and *in silico* evaluation steps (Figure 5); peptide sequences are mutated and then evaluated for predicted antimicrobial activity. The sequences with the greatest predicted activity are ranked highly and subjected to additional rounds of mutation and evaluation, until the predicted activity is maximized. Here, the authors used a linear regression model to predict which amino acid mutations would optimize the AMP's antimicrobial activity. Specifically, the regressor was trained on experimentally determined IC_{50} values for ∼180 peptides derived from a control peptide, temporin, against *E. coli*. With their model, the authors successfully optimized 44 peptides, which showed up to 160-fold improvement in their growth inhibitory activity against *E. coli in vitro*, relative to temporin.

Porto et al.[149] used an evolutionary algorithm to design AMPs by applying a fitness function that introduces new amino acids into a sequence by optimizing the ratio between hydrophobic moment and helix propensity. The authors generated 100 peptides from their chosen parent AMP, Pg-AMP1. For validation, they tested the antibacterial activity of 15 generated peptides against Pseudomonas *aeruginosa*, as well as the toxicity of the peptides against human blood cells. Interestingly, none of the 15 peptides were hemolytic, but eight of the 15 displayed antibacterial activity with an MIC less than or equal to their control peptide, magainin 2.

Torres et al.[150] developed a scoring function to mine the human proteome for AMPs. Their function assigned a score to each human proteome-encoded peptide based on their physicochemical properties—namely, sequence length, charge, and hydrophobicity. This led to the selection of desirable peptides with potential antimicrobial activity. The authors identified ∼2600 potential AMPs among 43,000 candidate peptides across a variety of organ systems (cardiovascular, nervous, renal, hematopoietic, and digestive). For validation, they synthesized 55 peptides and tested them against an array of five pathogenic bacterial species (*E. coli*, *P. aeruginosa*, *S. aureus*, *Klebsiella pneumoniae*, and *A. baumannii)*—35 peptides were active. Remarkably, further studies of their two most potent AMPs showed continued activity against drug-resistant *A. baumannii*, whereas polymyxin B, their control peptide antibiotic, did not.

Early QSAR approaches demonstrated the value of DL in AMP discovery. Cherkasov et al.[151,152] developed a simple three-layer FFNN to design novel AMPs. Here, the authors computer-generated two training sets: set A (∼900 peptides) and set B (500 peptides). Set A was designed such that the peptides were similar in composition to known AMPs. Set B was developed using the amino acid compositions of the most active peptides from set A. Then, they collected (1) 44 physicochemical descriptors (e.g., charge and hydrophobic moment) for each peptide; and (2) antimicrobial activity against *P. aeruginosa* in the form of IC_{50} values. Their neural network jointly learned the structure and activity of each peptide in the training set and was then applied to identify AMPs from a test set of ∼99,000 peptides generated using the amino acid composition of peptides from training set B. To validate their findings, the authors synthesized 200 peptides of varying predicted activity levels and tested them against *P. aeruginosa*. Of the 50 peptides highly predicted to be active, 47 were confirmed and were also more active than their control peptide, Bac2A. All 50 tested peptides that were predicted to be inactive were, indeed, inactive.

Das et al.[153] used a VAE to generate broad-spectrum AMPs with low toxicity. The authors trained their model on a subset of the UniProt sequence database[154] consisting of peptides 25 amino acids or less in length (93,000 unlabeled sequences and 5000 known AMP sequences). With their trained VAE, the authors generated ∼90,000 peptide sequences, which they screened *in silico* using an RNN to predict antimicrobial activity, broad-spectrum potency, toxicity, and secondary structure. Their RNN was trained on annotations extracted from several databases, including satPDB, DBAASP, ToxinPred, and AMPEP. Twenty of their most promising predicted peptides were selected for synthesis and wet lab validation of antimicrobial activity against *S. aureus* and *E. coli*. Among these 20, five peptides exhibited an MIC against both pathogens. The authors further highlighted two peptides, YI12 and FK13, for their *in vivo* potency against multidrug-resistant *K. pneumoniae*, reduced propensity for resistance in *E. coli* (relative to the carbapenem antibiotic imipenem), and their efficacy in a mouse model of infection.

Integration of ML for MOA prediction

Though their use in antibiotic discovery may be somewhat more focused on chemical and sequence space exploration during early discovery—largely due to the existence of empirical data with which to train—ML approaches can be integrated into every step of the antibiotic development pipeline. Precisely determining a novel compound's MOA is a challenging process requiring a large suite of untargeted techniques and, often, nonobvious interpretations of the resulting systems-scale datasets. Indeed, contemporary methods to accelerate MOA elucidation, such as high-content image analysis[155–157] and omics-based approaches,[158–160] produce large, multidimensional data outputs that are challenging to fully leverage. Appropriately trained ML algorithms have powerful discriminative ability and are beginning to prove their potential in MOA elucidation.[28,161] For example, Godinez et al.[29] implemented DL-based approaches with high-content image analysis for MOA prediction. Here, the authors trained a CNN on a collection of ∼1700 images of MCF-7 cancer cells that were chemically perturbed with 37 compounds across 12 unique mechanisms of action.[162] Their trained model was able to accurately predict the MOAs of each compound at eight different concentrations, providing a simple proof-of-concept for the potential of DL algorithms in high-content image analysis for drug classification.

A similar approach can be applied to antibiotic discovery using bacterial cell micrographs for MOA prediction (Figure 5).[156,163] Indeed, Zoffmann et al.[164] used an RF model to develop an automated bacterial cytological profile image analysis pipeline for the purposes of antibiotic discovery and MOA elucidation. Here, alterations in bacterial cell envelope morphology, nucleoid morphology, and membrane integrity were visualized using fluorescence microscopy. The model takes as input a compound label and the corresponding phenotypic fingerprint that is generated from over 100 features extracted from the images (cell morphology, DNA morphology, membrane fluorescence intensity, DNA fluorescence intensity, and membrane integrity). The authors created an *E. coli* reference image training set of six chemically perturbed conditions (sub-MIC concentrations of colistin, doxycycline, ceftriaxone, globomycin, levofloxacin, and mecillinam) and a DMSO control condition. Their model outputs a similarity score relative to the reference image dataset, effectively comparing the MOA of a naïve input molecule to each compound in the reference set. Following training, the authors validated their model's ability to identify compounds with a high similarity to the reference compounds using a test set of seven antibiotics with MOAs similar to those in the reference set. Additionally, they showed the model's ability to predict the MOA of novel compounds using a test set of eight compounds with an unknown or dissimilar MOA to an expanded training set (which included triclosan, trimethoprim, MD3, and nitrofuran). Here, they synthesized five analogs of boronate—suggested to be an inhibitor of the fatty acid biosynthesis enzyme FabI—and their model predicted three of these compounds to be highly similar to triclosan, a known FabI inhibitor. Furthermore, the authors showed that their ML image analysis approach was sufficiently robust to be used against *A. baumannii* to characterize antibiotic-specific phenotypes, highlighting this approach's potential to be used across bacterial species.

Beyond image classification, investigators have also successfully integrated ML with experimental multi-omics data to develop models for MOA prediction.[165–167] As a concrete example, Yang et al.[165] leveraged a multitask elastic net, which solves multiple linear regression problems simultaneously, to understand the mechanisms underlying antibiotic lethality. The authors performed an antibiotic–metabolite chemical screen in *E. coli* with a set of 206 diverse carbon, nitrogen, phosphorus, and sulfur metabolites. For each condition, IC_{50} values were collected for three antibiotics (ampicillin, ciprofloxacin, and gentamicin) at 13 concentrations, yielding a dataset of over 20,000 values. These data were used to create network metabolic models of *E. coli* for each metabolite condition, to be used as input in the multitask elastic net. This ensured that the model learned the relationships between each antibiotic and the metabolic pathways involved in their activity. They validated their model by identifying 13 pathways associated with central carbon metabolism and nucleotide biosynthesis that were involved in the lethality mechanisms of ampicillin, ciprofloxacin, and gentamicin—a relationship which has previously been described. Interestingly, the authors also showed that conventional methods of pathway analysis that use hit cutoff values did not identify some novel pathways that were identified using their ML approach. Specifically, their model implicated purine biosynthesis in antibiotic lethality. Further wet lab investigation found adenine limitation caused by antibiotic exposure induces purine biosynthesis, which drives central carbon metabolism and respiration, damaging DNA through the production of toxic metabolic byproducts.

OUTLOOK

The rapid development of diverse ML methods has positioned us to change the paradigm of preclinical antibiotic discovery and development (Figure 5). In its current state, the antibiotic discovery pipeline is insufficient. However, ML approaches show great promise in increasing efficiency and decreasing the cost of new antibiotic discovery. Perhaps the most exciting prospect of contemporary ML and DL approaches—when appropriately trained—is their ability to efficiently identify antibacterial molecules across numerous desirable properties. Indeed, it is relatively easy to find molecules that are antibacterial, but it is challenging to find molecules that are suitable for use as clinical antibiotics. Rapidly identifying the most viable molecules for downstream optimization will accelerate antibiotic development by minimizing attrition rates throughout the development pipeline. Additionally, there is a growing interest in personalized medicine where patient care and treatment, as well as disease prediction and diagnosis, are tailored to the individual.[168] In the context of antibiotic discovery, this means the development of patient-specific and pathogen-specific antibiotics, as well as next-generation rapid diagnostics, to avoid disrupting the native microbiota and decreasing the rate of dissemination of resistance. Importantly, other areas of drug discovery have shown promising advancement toward these goals in personalized medicine. In anticancer development, for example, immunotherapies and target-specific compounds have already been approved for clinical use.[169] Moreover, we have recently seen the development of a variety of point-of-care diagnostics for COVID-19.[170] The time is ripe for the widespread adoption of ML and DL approaches for new antibiotic discovery to increase the rate of new drug discovery and decrease the burden of the inevitable evolution of resistance. We posit that outpacing resistance is possible.

We emphasize that interdisciplinary collaboration is essential since both experimental data and ML model development must be robust to make accurate predictions in unexplored chemical/sequence spaces. Similarly, increased democratization of ML and DL resources for new antibiotic discovery will be instrumental in advancing antibiotic research toward our collective goals. ML approaches and tools are becoming widely accessible[171,172] even to those without strong computational expertise. However, unlike some fields that have large and reasonably well-controlled publicly available datasets, antibacterial research is somewhat lacking in the quantity and methodological transparency of easily accessible data. Open access to proprietary datasets[173] and the development of consortia[174] to allow for the collection of standardized antibacterial screening datasets for public use will increase the probability of the success of ML approaches in new antibiotic discovery. This should be a short-term priority for the field that can position us for long-term success.

AUTHOR CONTRIBUTIONS

T.L. and J.M.S. wrote and edited the manuscript.

ACKNOWLEDGMENTS

This paper was supported by the David Braley Centre for Antibiotic Discovery, the Weston Family Foundation, and the Natural Sciences and Engineering Research Council of Canada.

COMPETING INTERESTS

J.M.S. is cofounder and scientific director of Phare Bio.

REFERENCES

1. Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., Mcmanigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet*, *399*, 629–655.
2. *Antimicrobial resistance: Tackling a crisis for the health and wealth of nations*. (2014). Review on Antimicrobial Resistance.
3. *2021 Antibacterial agents in clinical and preclinical development: An overview and analysis*. (2022). Geneva: World Health Organization.
4. Shlaes, D. M. (2020). Antibacterial drugs: The last frontier. *ACS Infectious Diseases*, *6*, 1313–1314.
5. Thomas, D., & Wessel, C. (2022). *The state of innovation in antibacterial therapeutics*. Biotechnology Innovation Organization.
6. Plackett, B. (2020). Why big pharma has abandoned antibiotics. *Nature*, *586*, S50–S52.
7. Clancy, C. J., & Nguyen, M. H. (2020). Buying time: The AMR Action Fund and the state of antibiotic development in the United States 2020. *Open Forum Infectious Diseases*, *7*, ofaa464.
8. Hunter, P. (2020). A war of attrition against antibiotic resistance. *EMBO Reports*, *21*, e50807.
9. Brown, E. D., & Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature*, *529*, 336–343.
10. Lewis, K. (2020). The science of antibiotic discovery. *Cell*, *181*, 29–45.
11. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or genomical? *PLOS Biology*, *13*, e1002195.
12. Miethke, M., Pieroni, M., Weber, T., Brönstrup, M., Hammann, P., Halby, L., Arimondo, P. B., Glaser, P., Aigle, B., Bode, H. B., Moreira, R., Li, Y., Luzhetskyy, A., Medema, M. H., Pernodet, J.-L., Stadler, M., Tormo, J. R. N., Genilloud, O., Truman, A. W., … Müller, R. (2021). Towards the sustainable discovery and development of new antibiotics. *Nature Reviews Chemistry*, *5*, 726–749.
13. Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of Big Data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 69–80.
14. French, S., Guo, A. B. Y., & Brown, E. D. (2020). A comprehensive guide to dynamic analysis of microbial gene expression using the 3D-printed PFIbox and a fluorescent reporter library. *Nature Protocols*, *15*, 575–603.
15. French, S., Mangat, C., Bharat, A., Côté, J.-P., Mori, H., & Brown, E. D. (2016). A robust platform for chemical genomics in bacterial systems. *Molecular Biology of the Cell*, *27*, 1015–1025.
16. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, *2*, 2006. 0008.
17. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., … Aspuru-Guzik, A. N. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, *37*, 1038–1040.
18. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., Macnair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., … Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, *181*, 475–483.
19. Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., Dimaio, F., Carter, L., Chow, C. M., Montelione, G. T., & Baker, D. (2021). De novo protein design by deep network hallucination. *Nature*, *600*, 547–552.
20. Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature*, *557*, S55–S57.
21. Mak, K.-K., & Pichika, M. R. (2019). Artificial intelligence in drug development: Present status and future prospects. *Drug Discovery Today*, *24*, 773–780.
22. Dobson, C. M. (2004). Chemical space and biology. *Nature*, *432*, 824–828.
23. Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for Big Data: A review. *Big Data Research*, *2*, 87–93.
24. Tommasi, R., Brown, D. G., Walkup, G. K., Manchester, J. I., & Miller, A. A. (2015). ESKAPEing the labyrinth of antibacterial discovery. *Nature Reviews Drug Discovery*, *14*, 529–542.
25. Payne, D. J., Gwynn, M. N., Holmes, D. J., & Pompliano, D. L. (2007). Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nature Reviews Drug Discovery*, *6*, 29–40.
26. Li, L., Le, X., Wang, L., Gu, Q., Zhou, H., & Xu, J. (2015). Discovering new DNA gyrase inhibitors using machine learning approaches. *RSC Advances*, *5*, 105600–105608.
27. Mansbach, R. A., Leus, I. V., Mehla, J., Lopez, C. A., Walker, J. K., Rybenkov, V. V., Hengartner, N. W., Zgurskaya, H. I., & Gnanakaran, S. (2020). Machine learning algorithm identifies an antibiotic vocabulary for permeating Gram-negative bacteria. *Journal of Chemical Information and Modeling*, *60*, 2838–2847.
28. Santiago, M., Lee, W., Fayad, A. A., Coe, K. A., Rajagopal, M., Do, T., Hennessen, F., Srisuknimit, V., Müller, R., Meredith, T. C., & Walker, S. (2018). Genome-wide mutant profiling predicts the mechanism of a Lipid II binding antibiotic. *Nature Chemical Biology*, *14*, 601–608.
29. Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., & Zhang, X. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, *33*, 2010–2019.
30. Madhukar, N. S., Khade, P. K., Huang, L., Gayvert, K., Galletti, G., Stogniew, M., Allen, J. E., Giannakakou, P., & Elemento, O. (2019). A Bayesian machine learning approach for drug target identification using diverse data types. *Nature Communications*, *10*, 5221.
31. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*, 40–55.
32. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.
33. Zhu, H. (2020). Big Data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology*, *60*, 573–589.
34. David, L., Thakkar, A., Mercado, R. O., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: A review and practical guide. *Journal of Cheminformatics*, *12*, 56.

35. Liu, C., Hogan, A. M., Sturm, H., Khan, M. W., Islam, M. M., Rahman, A. S. M. Z, Davis, R., Cardona, S. T., & Hu, P. (2022). Deep learning-driven prediction of drug mechanism of action from large-scale chemical–genetic interaction profiles. *Journal of Cheminformatics*, *14*, 1–17.
36. Richter, M. F., Drown, B. S., Riley, A. P., Garcia, A., Shirai, T., Svec, R. L., & Hergenrother, P. J. (2017). Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature*, *545*, 299–304.
37. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*, 463–477.
38. Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, *16*, 687–694.
39. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*, 321–332.
40. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*, 12–18.
41. Kan, A. (2017). Machine learning applications in cell image analysis. *Immunology and Cell Biology*, *95*, 525–530.
42. Liu, Z., Deng, D., Lu, H., Sun, J., Lv, L., Li, S., Peng, G., Ma, X., Li, J., Li, Z., Rong, T., & Wang, G. (2020). Evaluation of machine learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences. *Frontiers in Microbiology*, *11*, 48.
43. Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *Journal of Machine Learning Research*, *22*(201), 1–7.
44. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*, *23*, 1241–1250.
45. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
46. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.
47. Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*, 239–263.
48. Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, *4*, 320–330.
49. Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods*, *13*, 603–604.
50. Hie, B., Bryson, B. D., & Berger, B. (2020). Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Systems*, *11*, 461–477.e9.
51. Da Silva, T. H., Hachigian, T. Z., Lee, J., & King, M. D. (2022). Using computers to ESKAPE the antibiotic resistance crisis. *Drug Discovery Today*, *27*, 456–470.
52. Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, *83*, 770–803.
53. Hong, J. (2011). Natural product diversity and its role in chemical biology and drug discovery. *Current Opinion in Chemical Biology*, *15*, 350–354.
54. Walsh, C. (2003). Where will new antibiotics come from? *Nature Reviews Microbiology*, *1*, 65–70.
55. Scherlach, K., & Hertweck, C. (2021). Mining and unearthing hidden biosynthetic potential. *Nature Communications*, *12*, 3864.
56. Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., … Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, *417*, 141–147.
57. Ōmura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., Kikuchi, H., Shiba, T., Sakaki, Y., & Hattori, M. (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 12215–12220.
58. Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D. H., & Wohlleben, W. (2009). CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *Journal of Biotechnology*, *140*, 13–17.
59. Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0—A versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research*, *41*, W204–W212.
60. Li, M. H., Ung, P. M., Zajkowski, J., Garneau-Tsodikova, S., & Sherman, D. H. (2009). Automated genome mining for natural products. *BMC Bioinformatics*, *10*, 1–10.
61. Skinnider, M. A., Johnston, C. W., Edgar, R. E., Dejong, C. A., Merwin, N. J., Rees, P. N., & Magarvey, N. A. (2016). Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, E6343–E6351.
62. Behsaz, B., Bode, E., Gurevich, A., Shi, Y.-N., Grundmann, F., Acharya, D., Caraballo-Rodríguez, A. M., Bouslimani, A., Panitchpakdi, M., Linck, A., Guan, C., Oh, J., Dorrestein, P. C., Bode, H. B., Pevzner, P. A., & Mohimani, H. (2021). Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. *Nature Communications*, *12*, 3225.
63. Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J. J., van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S. L., Lund, G., Epstein, S. C., Sisto, A. C., Charkoudian, L. K., Collemare, J., Linington, R. G., … Medema, M. H. (2020). MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*, *48*, D454–D458.
64. Chen, Y., Garcia De Lomana, M., Friedrich, N.-O., & Kirchmair, J. (2018). Characterization of the chemical space of known and readily obtainable natural products. *Journal of Chemical Information and Modeling*, *58*, 1518–1532.
65. Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G., & Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, *158*, 412–421.
66. Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., Clardy, J., Linington, R. G., & Fischbach, M. A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, *158*, 1402–1414.
67. Prihoda, D., Maritz, J. M., Klempir, O., Dzamba, D., Woelk, C. H., Hazuda, D. J., Bitton, D. A., & Hannigan, G. D. (2021). The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Natural Product Reports*, *38*, 1100–1108.
68. Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., & Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, *47*, e110.

69. Dias, T., Gaudêncio, S., & Pereira, F. (2019). A computer-driven approach to discover natural product leads for methicillin-resistant *Staphylococcus aureus* infection therapy. *Marine Drugs*, *17*, 16.
70. Walker, A. S., & Clardy, J. (2021). A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *Journal of Chemical Information and Modeling*, *61*, 2560–2571.
71. Liu, Z., Zhao, Y., Huang, C., & Luo, Y. (2021). Recent advances in silent gene cluster activation in Streptomyces. *Frontiers in Bioengineering and Biotechnology*, *9*, 632230.
72. De Avila E Silva, S., Echeverrigaray, S., & Gerhardt, G. J. L. (2011). BacPP: Bacterial promoter prediction—A tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of Theoretical Biology*, *287*, 92–99.
73. Kotopka, B J., & Smolke, C. D. (2020). Model-driven generation of artificial yeast promoters. *Nature Communications*, *11*, 2113.
74. Wu, M.- R., Nissim, L., Stupp, D., Pery, E., Binder-Nissim, A., Weisinger, K., Enghuus, C., Palacios, S. R., Humphrey, M., Zhang, Z., Maria Novoa, E., Kellis, M., Weiss, R., Rabkin, S. D., Tabach, Y., & Lu, T. K. (2019). A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nature Communications*, *10*, 2880.
75. Tian, J., Li, Q., Chu, X., & Wu, N. (2018). Presyncodon, a web server for gene design with the evolutionary information of the expression hosts. *International Journal of Molecular Sciences*, *19*, 3872.
76. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., & Church, G. M. (2021). Low-N protein engineering with data-efficient deep learning. *Nature Methods*, *18*, 389–396.
77. Von Nussbaum, F., Brands, M., Hinzen, B., Weigand, S., & Häbich, D. (2006). Antibacterial natural products in medicinal chemistry—Exodus or revival? *Angewandte Chemie*, *45*, 5072–5129.
78. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *46*, 3–26.
79. O'Shea, R., & Moser, H. E. (2008). Physicochemical properties of antibacterial compounds: Implications for drug discovery. *Journal of Medicinal Chemistry*, *51*, 2871–2878.
80. Zhao, S., Adamiak, J. W., Bonifay, V., Mehla, J., Zgurskaya, H. I., & Tan, D. S. (2020). Defining new chemical space for drug penetration into Gram-negative bacteria. *Nature Chemical Biology*, *16*, 1293–1302.
81. Farha, M. A., & Brown, E. D. (2015). Unconventional screening approaches for antibiotic discovery. *Annals of the New York Academy of Sciences*, *1354*, 54–66.
82. Gajdács, M. (2019). The concept of an ideal antibiotic: Implications for drug design. *Molecules*, *24*, 892.
83. Silver, L. L. (2016). A Gestalt approach to Gram-negative entry. *Bioorganic & Medicinal Chemistry*, *24*, 6379–6389.
84. Drew, K. L. M., Baiman, H., Khwaounjoo, P., Yu, B., & Reynisson, J. (2012). Size estimation of chemical space: How big is it? *Journal of Pharmacy and Pharmacology*, *64*, 490–495.
85. Walters, W. P., & Barzilay, R. (2021). Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, *54*, 263–270.
86. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*, 4977–5010.
87. Hajduk, P. J., & Greer, J. (2007). A decade of fragment-based drug design: Strategic advances and lessons learned. *Nature Reviews Drug Discovery*, *6*, 211–219.
88. Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, *194*, 178–180.
89. Klopman, G. (1984). Artificial intelligence approach to structure–activity studies. Computer automated structure evaluation of biological activity of organic molecules. *Journal of the American Chemical Society*, *106*, 7315–7321.
90. Ivanenkov, Y. A., Zhavoronkov, A., Yamidanov, R. S., Osterman, I. A., Sergiev, P. V., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Veselov, M. S., Ayginin, A. A., Kartsev, V. G., Skvortsov, D. A., Chemeris, A. V., Baimiev, A. K., Sofronova, A. A., Malyshev, A. S., Filkov, G. I., Bezrukov, D. S., Zagribelnyy, B. A., … Dontsova, O. A. (2019). Identification of novel antibacterials using machine learning techniques. *Frontiers in Pharmacology*, *10*, 913.
91. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, *28*, 2224–2232.
92. Kearnes, S., Mccloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, *30*, 595–608.
93. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, *59*, 3370–3388.
94. Sterling, T., & Irwin, J. J. (2015). ZINC 15 – Ligand discovery for everyone. *Journal of Chemical Information and Modeling*, *55*, 2324–2337.
95. Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. *Proceedings of the International Conference on Machine Learning*, *70*, 1945–1954.
96. Lim, J., Ryu, S., Kim, J. W., & Kim, W. Y. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, *10*, 1–9.
97. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, *4*, 268–276.
98. Bickerton, G. R, Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, *4*, 90–98.
99. Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. *Proceedings of the International Conference on Machine Learning*, *80*, 2323–2332.
100. Lin, E., Lin, C.-H., & Lane, H.-Y. (2020). Relevant applications of generative adversarial networks in drug design and discovery: Molecular de novo design, dimensionality reduction, and de novo peptide and protein design. *Molecules*, *25*, 3250.
101. Merk, D., Friedrich, L., Grisoni, F., & Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. *Molecular Informatics*, *37*, 1700153.
102. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, *14*, 3098–3104.
103. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., & Wichard, J. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Communications*, *11*, 10.
104. Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., … Golub, T. R. (2017). A next generation

connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, *171*, 1437–1452.e17.

105. Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, *4*, 120–131.

106. Jenkins, J. L., Bender, A., & Davies, J. W. (2006). In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technologies*, *3*, 413–421.

107. Gray, D. A., & Wenzel, M. (2020). Multitarget approaches against multiresistant superbugs. *ACS Infectious Diseases*, *6*, 1346–1365.

108. Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology*, *4*, 682–690.

109. Hase, T., Tanaka, H., Suzuki, Y., Nakagawa, S., & Kitano, H. (2009). Structure of protein interaction networks and their implications on drug design. *PLoS Computational Biology*, *5*, e1000550.

110. Nidhi, G. M., Davies, J. W., & Jenkins, J. L. (2006). Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *Journal of Chemical Information and Modeling*, *46*, 1124–1133.

111. Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R. P., & Pande, V. (2017). Is multitask deep learning practical for pharma? *Journal of Chemical Information and Modeling*, *57*, 2068–2076.

112. Liu, S., Qu, M., Zhang, Z., Cai, H., & Tang, J. (2022). Structured multi-task learning for molecular property prediction. *ArXiv220304695 Cs Q-Bio Stat*, *151*, 8906–8920.

113. Jin, W., Barzilay, R., & Jaakkola, T. (2020). Multi-objective molecule generation using interpretable substructures. *Proceedings of the International Conference on Machine Learning*, *119*, 4849–4859.

114. Khemchandani, Y., O'Hagan, S., Samanta, S., Swainston, N., Roberts, T. J., Bollegala, D., & Kell, D. B. (2020). DeepGraphMolGen, a multi-objective, computational strategy for generating molecules with desirable properties: A graph convolution and reinforcement learning approach. *Journal of Cheminformatics*, *12*, 1–17.

115. Ekins, S., Honeycutt, J. D, & Metz, J. T. (2010). Evolving molecules using multi-objective optimization: Applying to ADME/Tox. *Drug Discovery Today*, *15*, 451–460.

116. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, *44*, D1045–D1053.

117. Theuretzbacher, U., Outterson, K., Engel, A., & Karlén, A. (2020). The global preclinical antibacterial pipeline. *Nature Reviews Microbiology*, *18*, 275–285.

118. Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G., & Alon, U. (2006). A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, *3*, 623–628.

119. Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., Chun-Rong, L., Guenthner, D., Bovee, D., Olson, M. V., & Manoil, C. (2003). Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 14339–14344.

120. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D., & Palsson, B. O. (2020). A biochemically-interpretable machine learning classifier for microbial GWAS. *Nature Communications*, *11*, 2580.

121. Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Computational Biology*, *15*, e1007084.

122. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, *173*, 1581–1592.

123. Woo, G., Fernandez, M., Hsing, M., Lack, N. A., Cavga, A. D., & Cherkasov, A. (2020). DeepCOP: Deep learning-based approach to predict gene regulating effects of small molecules. *Bioinformatics*, *36*, 813–818.

124. Cao, X., Boyaci, H., Chen, J., Bao, Y., Landick, R., & Campbell, E. A. (2022). Basis of narrow-spectrum activity of fidaxomicin on *Clostridioides difficile*. *Nature*, *604*, 541–545.

125. Eyal, Z., Matzov, D., Krupkin, M., Wekselman, I., Paukner, S., Zimmerman, E., Rozenberg, H., Bashan, A., & Yonath, A. (2015). Structural insights into species-specific features of the ribosome from the pathogen *Staphylococcus aureus*. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E5805–E5814.

126. Brochado, A. R., Telzerow, A., Bobonis, J., Banzhaf, M., Mateus, A., Selkrig, J., Huth, E., Bassler, S., Zamarreño Beas, J., Zietek, M., Ng, N., Foerster, S., Ezraty, B., Py, B., Barras, F., Savitski, M. M., Bork, P., Göttig, S., & Typas, A. (2018). Species-specific activity of antibacterial drug combinations. *Nature*, *559*, 259–263.

127. Shapiro, J. A., Kaplan, A. R., & Wuest, W. M. (2019). From general to specific: Can *Pseudomonas* primary metabolism be exploited for narrow-spectrum antibiotics? *Journal of Chemical Biology*, *20*, 34–39.

128. Macnair, C. R., Tsai, C. N., & Brown, E. D. (2020). Creative targeting of the Gram-negative outer membrane in antibiotic discovery. *Annals of the New York Academy of Sciences*, *1459*, 69–85.

129. Keith, C. T., Borisy, A. A., & Stockwell, B. R. (2005). Multicomponent therapeutics for networked systems. *Nature Reviews Drug Discovery*, *4*, 71–78.

130. Hall, M. J., Middleton, R. F., & Westmacott, D. (1983). The fractional inhibitory concentration (FIC) index as a measure of synergy. *Journal of Antimicrobial Chemotherapy*, *11*, 427–433.

131. Baeder, D. Y., Yu, G., Hozé, N., Rolff, J., & Regoes, R. R. (2016). Antimicrobial combinations: Bliss independence and Loewe additivity derived from mechanistic multi-hit models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150294.

132. Tyers, M., & Wright, G. D. (2019). Drug combinations: A strategy to extend the life of antibiotics in the 21st century. *Nature Reviews Microbiology*, *17*, 141–155.

133. Chandrasekaran, S., Cokol-Cakmak, M., Sahin, N., Yilancioglu, K., Kazan, H., Collins, J. J., & Cokol, M. (2016). Chemogenomics and orthology-based design of antibiotic combination therapies. *Molecular Systems Biology*, *12*, 872.

134. Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K. M., Lee, K. J., Wong, A., Shales, M., Lovett, S., Winkler, M. E., Krogan, N. J., Typas, A., & Gross, C. A. (2011). Phenotypic landscape of a bacterial cell. *Cell*, *144*, 143–156.

135. Mason, D. J., Stott, I., Ashenden, S., Weinstein, Z. B., Karakoc, I., Meral, S., Kuru, N., Bender, A., & Cokol, M. (2017). Prediction of antibiotic interactions using descriptors derived from molecular structure. *Journal of Medicinal Chemistry*, *60*, 3902–3912.

136. Wildenhain, J., Spitzer, M., Dolma, S., Jarvik, N., White, R., Roy, M., Griffiths, E., Bellows, D. S., Wright, G. D., & Tyers, M. (2015). Prediction of compound synergism from chemical–genetic interactions by machine learning. *Cell Systems*, *1*, 383–395.

137. Weinstein, Z. B., Bender, A., & Cokol, M. (2017). Prediction of synergistic drug combinations. *Current Opinion in Systems Biology*, *4*, 24–28.

138. Cokol, M., Li, C., & Chandrasekaran, S. (2018). Chemogenomic model identifies synergistic drug combinations robust to the pathogen microenvironment. *PLoS Computational Biology*, *14*, e1006677.

139. Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., & Deng, L. (2019). DrugCombDB: A comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Research*, *48*, D871–D881.

140. Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., Nguyen, T., Zaslavskiy, M., Jang, I. S., Ghazoui, Z., Ahsen, M. E., Vogel, R., Neto, E. C., Norman, T., Tang, E. K. Y., … Saez-Rodriguez, J. (2019). Community assessment to

advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*, *10*, 2674.

141. Preuer, K., Lewis, R. P. I., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). DeepSynergy: Predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, *34*, 1538–1546.
142. O'neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., Arthur, W., Cristescu, R., Haines, B. B., Winter, C., Zhang, T., Bloecher, A., & Shumway, S. D. (2016). An unbiased oncology compound screen to identify novel combination strategies. *Molecular Cancer Therapeutics*, *15*, 1155–1162.
143. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., Van Dyk, E., Chang, H., De Silva, H., Heyn, H., Deng, X., Egan, R K., Liu, Q., … Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, *166*, 740–754.
144. Kuru, H. I., Tastan, O., & Cicek, E. (2021). MatchMaker: A deep learning framework for drug synergy prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatic*, *19*, 2334–2344.
145. Liu, Q., & Xie, L. (2021). TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS Computational Biology*, *17*, e1008653.
146. Jin, W., Stokes, J. M., Eastman, R. T., Itkin, Z., Zakharov, A. V., Collins, J. J., Jaakkola, T. S., & Barzilay, R. (2021). Deep learning identifies synergistic drug combinations for treating COVID-19. *Proceedings of the National Academy of Sciences of the United States of America*, *118*, e2105070118.
147. Mahlapuu, M., Håkansson, J., Ringstad, L., & Björn, C. (2016). Antimicrobial peptides: An emerging category of therapeutic agents. *Frontiers in Cellular and Infection Microbiology*, *6*, 194.
148. Yoshida, M., Hinkley, T., Tsuda, S., Abul-Haija, Y. M., Mcburney, R. T., Kulikov, V., Mathieson, J. S., Garcia Reyes, S., Castro, M. D., & Cronin, L. (2018). Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem*, *4*, 533–543.
149. Porto, W. F., Irazazabal, L., Alves, E. S. F., Ribeiro, S. M., Matos, C. O., Pires, Ã. S., Fensterseifer, I. C. M., Miranda, V. J., Haney, E. F., Humblot, V., Torres, M. D. T., Hancock, R. E. W., Liao, L. M., Ladram, A., Lu, T. K., De La Fuente-Nunez, C., & Franco, O. L. (2018). In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nature Communications*, *9*, 1490.
150. Torres, M. D. T., Melo, M. C. R., Crescenzi, O., Notomista, E., & De La Fuente-Nunez, C. (2022). Mining for encrypted peptide antibiotics in the human proteome. *Nature Biomedical Engineering*, *6*, 67–75.
151. Cherkasov, A., Hilpert, K., Jenssen, H., Fjell, C. D., Waldbrook, M., Mullaly, S. C., Volkmer, R., & Hancock, R. E. W. (2009). Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chemical Biology*, *4*, 65–74.
152. Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A., Panté, N., Hancock, R. E. W., & Cherkasov, A. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of Medicinal Chemistry*, *52*, 2006–2015.
153. Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Dos Santos, C., Chen, P.-Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., & Mojsilovic, A. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, *5*, 613–623.
154. The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*, D480–D489.
155. Peach, K. C., Bray, W. M., Winslow, D., Linington, P. F., & Linington, R. G. (2013). Mechanism of action-based classification of antibiotics using high-content bacterial image analysis. *Molecular BioSystems*, *9*, 1837–1848.
156. Nonejuie, P., Burkart, M., Pogliano, K., & Pogliano, J. (2013). Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 16169–16174.
157. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., & Carpenter, A. E. (2021). Image-based profiling for drug discovery: Due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, *20*, 145–159.
158. Zampieri, M., Szappanos, B., Buchieri, M. V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., Papp, B., & Sauer, U. (2018). High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Science Translational Medicine*, *10*, eaal3973.
159. Wacker, S. A., Houghtaling, B. R., Elemento, O., & Kapoor, T. M. (2012). Using transcriptome sequencing to identify mechanisms of drug action and resistance. *Nature Chemical Biology*, *8*, 235–237.
160. Ribeiro Da Cunha, B., Fonseca, L. S. P., & Calado, C. L. R. C. (2020). Metabolic fingerprinting with Fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation. *Metabolites*, *10*, 145.
161. Santa Maria, J. P., Park, Y., Yang, L., Murgolo, N., Altman, M. D., Zuck, P., Adam, G., Chamberlin, C., Saradjian, P., Dandliker, P., Boshoff, H. I. M., Barry, C. E., Garlisi, C., Olsen, D. B., Young, K., Glick, M., Nickbarg, E., & Kutchukian, P. S. (2017). Linking high-throughput screens to identify MoAs and novel inhibitors of *Mycobacterium tuberculosis* dihydrofolate reductase. *ACS Chemical Biology*, *12*, 2448–2456.
162. Ljosa, V., Sokolnicki, K. L., & Carpenter, A. E. (2012). Annotated high-throughput microscopy image sets for validation. *Nature Methods*, *9*, 637–637.
163. Quach, D. T., Sakoulas, G., Nizet, V., Pogliano, J., & Pogliano, K. (2016). Bacterial cytological profiling (BCP) as a rapid and accurate antimicrobial susceptibility testing method for *Staphylococcus aureus*. *EBioMedicine*, *4*, 95–103.
164. Zoffmann, S., Vercruysse, M., Benmansour, F., Maunz, A., Wolf, L., Blum Marti, R., Heckel, T., Ding, H., Truong, H. H., Prummer, M., Schmucki, R., Mason, C. S., Bradley, K., Jacob, A. I., Lerner, C., Araujo Del Rosario, A., Burcin, M., Amrein, K. E., & Prunotto, M. (2019). Machine learning-powered antibiotics phenotypic drug discovery. *Scientific Reports*, *9*, 5013.
165. Yang, J. H., Wright, S. N., Hamblin, M., Mccloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C., & Collins, J. J. (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, *177*, 1649–1661.e9.
166. Ribeiro Da Cunha, B., Fonseca, L. S. P., & Calado, C. L. R. C. (2021). Simultaneous elucidation of antibiotic mechanism of action and potency with high-throughput Fourier-transform infrared (FTIR) spectroscopy and machine learning. *Applied Microbiology and Biotechnology*, *105*, 1269–1286.
167. Patel-Murray, N. L., Adam, M., Huynh, N., Wassie, B. T., Milani, P., & Fraenkel, E. (2020). A multi-omics interpretable machine learning model reveals modes of action of small molecules. *Scientific Reports*, *10*, 954.
168. Dugger, S. A., Platt, A., & Goldstein, D. B. (2018). Drug development in the era of precision medicine. *Nature Reviews Drug Discovery*, *17*, 183–196.
169. Döhner, H., Wei, A. H., & Löwenberg, B. (2021). Towards precision medicine for AML. *Nature Reviews Clinical Oncology*, *18*, 577–590.
170. Weissleder, R., Lee, H., Ko, J., & Pittet, M. J. (2020). COVID-19 diagnostics in context. *Science Translational Medicine*, *12*, eabc1931.

171. Melo, M. C. R., Maasch, J. R. M. A., & De La Fuente-Nunez, C. (2021). Accelerating antibiotic discovery through artificial intelligence. *Communications Biology*, *4*, 1–13.
172. Jukič, M., & Bren, U. (2022). Machine learning in antibacterial drug design. *Frontiers in Pharmacology*, *13*, 864412.
173. Kim, W., Krause, K., Zimmerman, Z., & Outterson, K. (2021). Improving data sharing to increase the efficiency of antibiotic R&D. *Nature Reviews Drug Discovery*, *20*, 1–2.
174. Ackloo, S., Al-Awar, R., Amaro, R. E., Arrowsmith, C. H., Azevedo, H., Batey, R. A., Bengio, Y., Betz, U. A. K., Bologa, C. G., Chodera, J. D., Cornell, W. D., Dunham, I., Ecker, G. F., Edfeldt, K., Edwards, A. M., Gilson, M. K., Gordijo, C. R., Hessler, G., … Willson, T. M. (2022). CACHE (Critical Assessment of Computational Hit-finding Experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry*, *6*, 287–295.

How to cite this article: Lluka, T., & Stokes, J. M. (2022). Antibiotic discovery in the artificial intelligence era. *Ann NY Acad Sci*., 1–20. https://doi.org/10.1111/nyas.14930.