REVIEW

Leveraging artificial intelligence in the fight against infectious diseases

Felix Wong[1,2], Cesar de la Fuente-Nunez[3,4,5]*, James J. Collins[1,2,6]*

Despite advances in molecular biology, genetics, computation, and medicinal chemistry, infectious disease remains an ominous threat to public health. Addressing the challenges posed by pathogen outbreaks, pandemics, and antimicrobial resistance will require concerted interdisciplinary efforts. In conjunction with systems and synthetic biology, artificial intelligence (AI) is now leading to rapid progress, expanding anti-infective drug discovery, enhancing our understanding of infection biology, and accelerating the development of diagnostics. In this Review, we discuss approaches for detecting, treating, and understanding infectious diseases, underscoring the progress supported by AI in each case. We suggest future applications of AI and how it might be harnessed to help control infectious disease outbreaks and pandemics.

Infectious diseases, caused by transmissible pathogens including bacteria, eukaryotes, and viruses, continue to challenge scientists and clinicians despite advances in medicine and basic research over the past few decades. Limitations to the fast and accurate detection of infections, as well as expanding antimicrobial resistance, exacerbate these challenges (Box 1). Basic research has aimed to expand our knowledge and provide solutions, including development of anti-infective therapies, preventative measures, and fast and accurate diagnostic tools. In particular, systems and synthetic biology approaches have led to biotechnological and medical innovations—including drug treatments and modalities, vaccines, and diagnostics—that have improved how we deal with infectious diseases.

The fields of systems and synthetic biology emerged from two key developments: (i) the generation and synthesis of quantitative biological hypotheses and data from wet-lab experiments, sequencing, and systems-level modeling; and (ii) an understanding of the modularity and programmability of nucleic acids, peptides, and other biomolecules, which enables control of biology. Artificial intelligence (AI), which focuses on developing machines capable of reasoning with data, has recently matured into an exciting field that draws on both these features to accelerate scientific discovery. Because AI-based approaches can integrate large amounts of quantitative and omics data, they are particularly adept at dealing with biological complexity, extending our knowledge and facilitating our efforts to reverse engineer and control biology. AI-based approaches are particularly useful in addressing the problem of infectious diseases, which are complex across different scales, ranging from cells to communities, and for which advances in medicine and biotechnology are essential drivers of progress. In this Review, we discuss major areas in which AI-based approaches, applied to systems and synthetic biology, are substantively empowering our research to fight infectious diseases.

"There is...an urgent need for new anti-infective treatments, particularly ones that represent unprecedented chemical spaces or therapeutic modalities."

AI for anti-infective drug discovery

Anti-infective drugs, comprising antibacterials, antivirals, antifungals, and antiparasitics, have become less effective treatments as a result of the spread of drug resistance. There is therefore an urgent need for new anti-infective treatments, particularly ones that represent unprecedented chemical spaces or therapeutic modalities. AI, and in particular machine learning (ML), a subfield of AI that uses data to train machines to make predictions, has foremost been helpful in facilitating searches of small-molecule databases, such as the ZINC15 (*1*). ML approaches to anti-infective drug discovery have centered on training models to identify new drugs or new uses of existing drugs (Fig. 1). As the number of drug-like small molecules is essentially infinite (as large as ~10^{60} (*2*), and possibly larger, given that typical antibiotics may not be traditionally drug-like (*3*)], a major benefit of ML approaches is that they can virtually screen compound libraries at a scale (>10^9 compounds) that would be impossible to screen empirically.

Anti-infective drug discovery has benefited particularly from AI integration for several reasons. First, in contrast to cancer or other diseases in which mechanism-driven approaches have remained dominant, infectious diseases are generally phenotype-driven; that is, these diseases proceed from the physiological characteristics of infectious agents rather than from their genetic or molecular compositions. The discovery of some of the first widely used antibacterials, antivirals, antiparasitics, and antifungals stemmed from observations of their inhibitory effects against pathogens or the symptoms caused by infections. This phenotypic line of discovery is as relevant today as it was decades ago, especially as innovations in high-throughput screening and the design of chemical libraries have enabled more quantitative and customizable discovery efforts. The focus on phenotypes implies that drug polypharmacologic effects can be common to anti-infective drugs and that biological information can be integrated across different macromolecular drug targets (*4*). Phenotypic properties are well suited for analysis by ML because ML can both unify and disentangle the different types of biological information that impinge on these readouts. Second, most anti-infective drugs are small molecules, whose chemical structures can be modeled computationally as graphs comprising vertices and edges, and additional programmable modalities, including target-binding nucleic acids called aptamers (*5*) and antimicrobial peptides (AMPs) (*6*–*8*), are currently in development. Supervised graph neural networks (*9*–*11*), unsupervised generative models (*12*, *13*)—which are ML models capable of producing outputs similar to their training data—and other recent advances in ML architectures (Box 1) enable computers to learn, predict, or design patterns in chemical structures, offering powerful tools for modeling small molecules. The use of ML to make biologically relevant predictions from sequences of nucleic acids or amino acids allows for ML-guided design relevant to these therapeutic modalities, as exemplified by protein structure prediction platforms such as AlphaFold and RoseTTAFold (*14*, *15*). Lastly, infectious diseases are typically caused by pathogens that are, or can be, well characterized. This biological tractability contrasts with complex diseases such as neurodegeneration, for which our incomplete mechanistic understanding remains a major bottleneck. Our clearer understanding and larger databases (*16*–*18*) of the gene and

[1]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [2]Institute for Medical Engineering and Science and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [4]Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104, USA. [5]Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA 19104, USA. [6]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA.
*Corresponding author. Email: cfuente@upenn.edu (C.F.N.); jimjc@mit.edu (J.J.C.)

protein networks of bacteria, viruses, and even simple eukaryotes—as compared with human cell types—may allow ML-driven approaches to make more-accurate predictions and better identify drug mechanisms of action (MoAs) (*19*–*21*).

Despite these advantages, there are outstanding issues in applying ML, and AI more broadly, to anti-infective drug discovery. One major challenge is that it is unclear how well ML models generalize to unexplored biomolecular spaces. For instance, we have previously screened a library of small molecules for growth inhibitory activity against *Escherichia coli* and used this phenotypic information to train graph neural networks to predict the antibiotic activities of small molecules—including halicin—on the basis of their chemical structures (*9*). Yet these models performed best at predicting compounds in well-known antibiotic classes, such as β-lactams and quinolones. To tap into previously unexplored search spaces, different approaches are needed. For example, a suboptimal solution was implemented during the course of a genetic algorithm—an algorithm that iteratively evolves its inputs to optimize a property—to identify the synthetic peptide guavanin 2. This peptide was subsequently synthesized and effectively killed bacteria in a preclinical mouse model, suggesting that the model could generalize at the cost of optimality (*22*). Recently, emerging computational approaches have made it possible to also mine proteomes for antibiotic discovery, leading to the identification of thousands of antimicrobials in both extant and extinct organisms (*6*, *23*).

Overall, lead molecules are only as structurally novel as the chemical spaces that are explored, and ML-driven approaches are limited by both the structural diversity of the training sets and the ability of model architectures to prioritize novelty. Organocatalysis and cascade reaction sequences, which are chemical synthesis methods that have recently opened up chemical spaces, can provide useful experimental starting points for generating structurally diverse small molecules (*24*). In contrast, the computational enumeration of all feasible small molecules containing atoms found in most drugs, as provided by the GDB datasets (*25*), presents opportunities to exhaustively sample chemical spaces of small molecules, with the caveat that other computational models are needed to accurately predict synthesizability. Nucleic acid–and peptide-encoded combinatorial libraries of small molecules (*26*) and peptides (*27*), as well as designable aptamers (*28*), can further extend search spaces of interest. In each case, generalizability is paramount to ML models. Improving generalizability will require the application of previously unexplored paradigms and models with improved inference capabilities, for instance, few-shot models, which are ML models that extrapolate from scarce training data (corresponding to undersampled regions of search spaces), or multitask models, which are ML models that combine information from diverse inputs. Models such as these will help to identify only the most promising drug candidates (*29*). Providing "negative" data (e.g., tested compounds that are not active) is also essential for ML model training and benchmarking, and when ML models are applied to challenging test sets, it is important that their limitations are clearly expressed (e.g., through confidence information). To express these limitations, interpretable or explainable ML approaches can be used to capture the specific aspects of training data that models have learned by pinpointing the input structural features (explainable ML) or the parts of the model that lead to a prediction (interpretable ML) (*30*).

Another key challenge in AI for anti-infective drug discovery is the need for improved mechanistic models to complement phenotypic approaches. Whereas ML models have been useful for identifying drug candidates on the basis of phenotypic information (*9*, *12*, *31*–*33*), more work is needed before models can accurately predict drug–target interactions and MoAs. These drug attributes remain important in light of antimicrobial resistance and the fact that we are still learning about the MoAs of anti-infective drugs discovered decades ago (*34*). Protein structure predictions (*14*, *15*) and other resources now provide structural information that can inform target-based predictions—although not knowing a protein's structure has not typically limited drug discovery (*35*). Recent studies have highlighted that improvements in molecular docking—which predicts binding affinities between ligands and targets on the basis of structural information—are still needed to accurately identify antibiotic MoAs, and that ML-driven approaches can improve prediction accuracy (*36*). Molecular docking approaches have largely focused on small-molecule ligands, but target predictability is just as important for AMPs, which often have nonspecific membrane-active MoAs (*22*, *31*, *32*), as well as aptamers. Improvements in target-centric approaches can facilitate the discovery of compounds with specific binding activity and lead to improved biological understanding, which can inform predictions of emergent properties such as drug interactions and synergies. Of particular relevance to antibiotic resistance, a better understanding of how compounds interact with membranes is crucial for discovering drugs that are active against Gram-negative bacteria, whose outer membranes have proven particularly difficult to penetrate (*37*).

Drug development is a lengthy and intricate process influenced by numerous factors such as safety, cost, manufacturing, and clinical trial outcomes. For

Box 1. Overarching challenges in infectious diseases and concepts in AI.

Pathogen outbreaks and pandemics: Recent outbreaks include COVID-19, mpox, Marburg virus, H5N1 influenza, Ebola, measles, Zika, *E. coli*, and MERS. Challenges include detecting outbreaks and new pathogens, understanding disease biology, and developing preventive measures.

Antimicrobial resistance and anti-infective drug discovery: Problematic pathogens include carbapenem-resistant Enterobacteriaceae (CRE), methicillin-resistant *Staphylococcus aureus* (MRSA), multidrug-resistant tuberculosis (MDR-TB), vancomycin-resistant Enterococcus (VRE), extended-spectrum beta-lactamase (ESBL)–producing bacteria, and drug-resistant *Candida auris*, *Neisseria gonorrhoeae*, *P. falciparum*, and *Toxoplasma gondii*. Challenges include practicing antimicrobial stewardship (the appropriate and responsible use of anti-infective drugs), developing new classes of anti-infective drugs, potentiating existing drugs against resistant infections, and understanding drug MoAs.

Neglected, persistent, and difficult-to-treat infections: Examples include neglected tropical diseases, chronic hepatitis B and C, chronic fungal infections, Lyme disease, infections in low-resource populations, and HIV/AIDS. Challenges include developing low-cost and field-deployable diagnostics, improving the accuracy of diagnostic tests, improving the detection of antimicrobial resistance, and making effective disease treatments available.

Artificial intelligence and machine learning: ML is a subfield of AI, and its approaches can be classified as supervised (model is told what property to predict), unsupervised (model is not told what property to predict), or reinforcement learning (model optimizes for feedback). Neural networks are a common ML architecture comprising interconnected layers of basic processing units (neurons). Different types of neural networks exist, including those that predict properties of graph-based inputs (graph neural networks), generate data by compressing what the model has learned (variational autoencoders), process sequential data (long short-term memory), and model complex dependencies by using attention mechanisms to focus on specific input elements (transformers). Not all models are neural networks, and simpler models include random forests (ensembles of decision trees), support vector machines (classifiers that separate data points on a plot), and regression models (functions that explicitly model the input–output relationship).

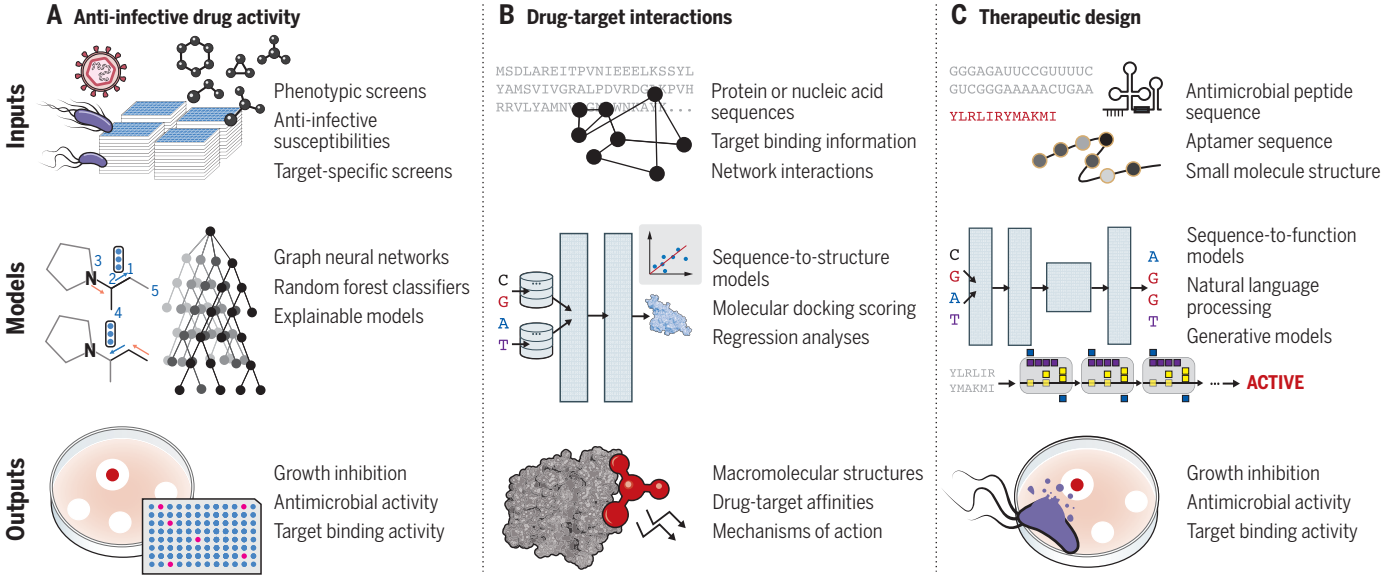

Fig. 1. AI can predict anti-infective drug activity, drug–target interactions, and therapeutic design. Examples of AI model inputs, model architectures or types, and model outputs relevant to anti-infective drug discovery include those focusing on drug activity (**A**), drug–target interactions and MoAs (**B**), and programmable therapeutic design (**C**). Inputs, models, and outputs shown are representative, in part, of those discussed in (*9–15*, *19–22*, *28–33*, *36*, *39*, *40*).

anti-infective drugs in particular, toxicity to host cells is a common liability. Drugs can be toxic in different ways (e.g., cytotoxic, hemolytic, and genotoxic), and ML models predicting toxicity have been limited by factors such as the lack of high-quality datasets (*38*). Absorption, distribution, metabolism, and excretion (ADME) properties, including chemical instability in solution and metabolic breakdown, are also needed to filter out drug candidates that are nonselective or unsuitable for medicinal use. Although high-throughput screens have focused on in vitro testing, there is substantial unmet need for anti-infective drugs that are effective against systemic infections. Predicting efficacy in animal models of acute systemic infections is a challenging task that has not yet been addressed by ML-driven approaches.

We anticipate that active areas to watch are those that combine experimental and computational approaches to address model predictive power and data scarcity. ML approaches that incorporate information from scarce training data, as well as more-extensive search spaces, are likely to substantively augment anti-infective drug discovery. To guide experimental methods to augment search spaces, generative ML models will continue to propose chemical structures and peptide sequences de novo that can be synthesized and evaluated. Generative platforms such as GPT-4 and NVIDIA's BioNeMo can also facilitate drug discovery by integrating disparate streams of scientific information to improve our understanding of the underlying biology and chemistry. Interpretable or explainable ML approaches (e.g., for graph neural networks) can offer powerful ways of inferring salient structural features or improving model learning from data. Computational pipelines that leverage structural predictions of proteins and other macromolecules provide a complementary way to improve model predictive power. Detailed molecular dynamics simulations and ML-augmented approaches to docking exemplify techniques that can better predict interactions between drugs and macromolecules (*36*, *39*). We anticipate that sequence-to-structure models, such as AlphaFold for proteins or FARFAR2 for RNAs (*14*, *40*), will also be useful for structure-guided design. Such models can be used to tune therapeutic candidates to achieve specific structures, bridging structural predictions with the productive augmentation of search spaces.

AI for infection biology and infection-related contexts

Bacterial, eukaryotic, and viral pathogens infect diverse hosts and trigger complex host responses. Pathogen load, host immunity, treatments administered, and other factors influence the course of infection. Supervised ML models have been used to analyze structured and unstructured nucleic acid, protein, glycan, and cellular phenotypic datasets to identify critical features and molecular networks involved in host–pathogen interactions and immune responses (Fig. 2) (*41–45*). Various supervised and unsupervised ML models, including random forest classifiers and complex language models (models designed to understand or generate text), have been applied to identify genes and protein–protein interactions associated with host cell changes, predict immunogenicity, and evaluate pathogen killing, host cell adaptation, and virulence. Additionally, supervised models have been used to guide the development of vaccines and therapeutic drugs through the optimization of gene expression and antigen prediction and selection (*46*, *47*). Reverse vaccinology, which bases antigen prediction on immunologic and genomic information, has been facilitated by supervised ML approaches, including Vaxign-ML (*47*).

In general, ML has made an outsized contribution to analyzing large and often convoluted datasets in infectious diseases research. Although these examples illustrate the promise of using ML to elucidate key factors underlying infections and how infections progress within hosts, understanding host–pathogen interactions and immune responses remains a challenging biological problem. This problem can be addressed by integrating high-throughput datasets—including sequencing, structural, and microscopy data—with detailed mechanistic studies, experimentation, and infection models. Mechanistic and experimental studies, however, are typically low-throughput, constraining the generalizability of AI-guided approaches that rely on them. Experiments in which large datasets are systematically acquired and analyzed across different infection contexts, for instance through comprehensive CRISPR screens, RNA sequencing, and mass spectrometry, would foster the development of AI models that extend beyond tools for data analysis and make generalizable hypotheses and inferences. Parameterizing these efforts with biological sequences or chemical structures, such as small molecules, guide RNAs, or amino acid sequences, would offer tunable approaches to investigating infection

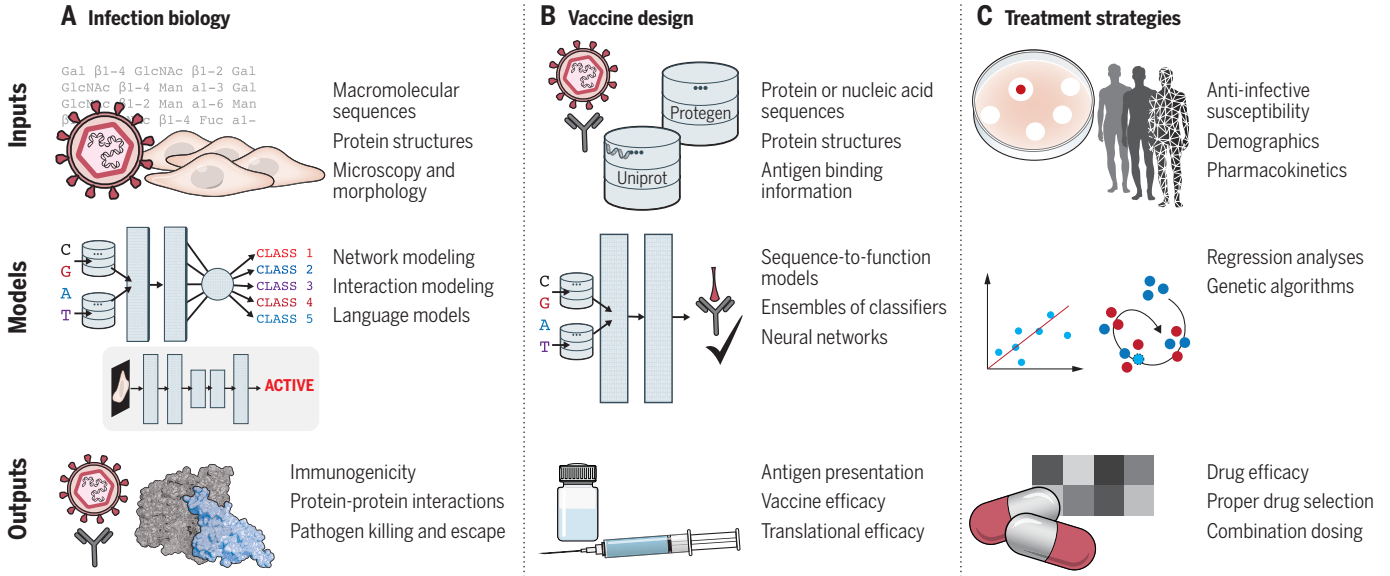

Fig. 2. AI can elucidate infection biology, facilitate vaccine design, and inform treatment strategies. Examples of AI model inputs, model architectures or types, and model outputs focusing on infection biology (**A**), vaccine design (**B**), and anti-infective drug treatment strategies (**C**). Inputs, models, and outputs shown are representative, in part, of those discussed in (*41*–*49*, *51*–*54*).

biology. In one example of a sequence-guided approach, a recent study developed unsupervised language models of influenza, HIV-1, and severe acute respiratory syndrome coronavirus 2 viral proteins based on amino acid sequence information and accurately predicted escape patterns that allow these pathogens to evade the human immune system (*45*). ML models that can make specific assumptions about biology, such as the relevance of syntax (grammar) and semantics (meaning) in biological sequences, or leverage structural information have the potential to guide the generation of biological hypotheses and improve generalizability.

Additionally, ML has productively processed microscopy datasets relevant to infection biology. Various forms of microscopy, including light and electron microscopy, have been used to generate datasets underlying ML models that detect bacteria, fungi, parasites, and viruses in host cells. These analyses have led to insights in host–pathogen biology, for instance by elucidating the developmental morphologies of *Plasmodium falciparum* in human red blood cells using multicolor fluorescence microscopy (*48*) and identifying virulence factors involved in *Mycobacterium abscessus* pathogenesis from high-content imaging and phenogenomic data (*49*).

Sequence-based ML approaches to messenger RNA and nucleic acid vaccines can accelerate design, and the turnaround times for the synthesis and experimental validation of these vaccines are short (*50*). Protein structure–based vaccine design (*51*) can also be augmented with computational predictions from AlphaFold or RoseTTAFold. Yet the use of ML for vaccine development faces several challenges, including poor data quality, limited data availability and generalizability, and complicated testing procedures. Limited or only low-quality data may be available for certain populations or diseases, particularly for neglected tropical diseases, and these limitations can influence the choices of target antigens and constrain ML models that predict antigen presentation and vaccine targets. Different infections have different host contexts, and ML models predicting the efficacy of vaccines, which modulate immunity in host cells, may be less generalizable to biological contexts than those for anti-infective drugs. Furthermore, the validation of vaccine candidates can be time-consuming and expensive, requiring delivery to host cells and suitable immunogenicity assays. To begin addressing these challenges, comprehensive benchmarking datasets for antigen selection and vaccine efficacy will be needed. These datasets will help to standardize data quality and improve the predictive power of next-generation ML approaches to vaccine development.

"Better leveraging of AI to address infectious diseases will require a collaborative effort among scientists, clinicians, and public health officials."

ML has also informed clinical decision-making in infection contexts. A recent study used regression models to implement personalized antibiotic recommendations that minimized the risk of urinary tract and wound infections (*52*). However, a general bottleneck in using ML to design treatment strategies is the need for data and models that are relevant to specific infection settings. An earlier study used support vector machines to analyze bacterial gene expression patterns in human patients, representing an important step toward showing that ML models can provide useful biological information relevant to clinical infections (*53*). Moving forward, multidimensional predictions of how anti-infective drugs and vaccines interact with model hosts and humans will help improve treatment strategies, anticipate adverse effects, and potentially increase success rates for new drugs in clinical trials.

As new datasets and models are needed to improve the application of ML to infection-related contexts, we anticipate that future work will make biology more "embeddable"—that is, able to be represented by low-dimensional features, such as sequences, vectors, or graphs. Integrating ML with next-generation systems and synthetic biology methods for cellular profiling will drive progress in this area. For instance, combining high-throughput screens and microscopy with precise methods for biological control, such as gene editing or optogenetics, would generate data relevant to key processes such as host cell stress responses, enabling manipulation of these pathways to address infectious diseases. In ML, promising types of language models include large language models, which are trained on large amounts of text data, and fine-tuned language

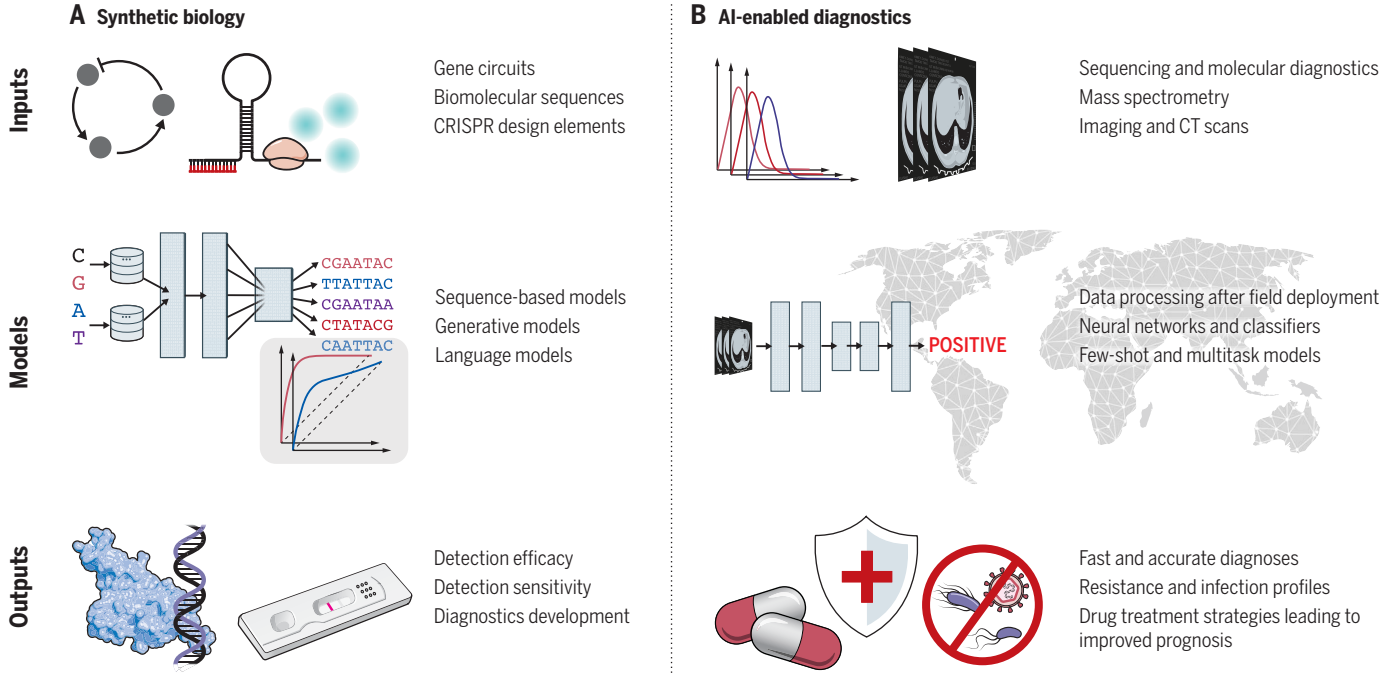

Fig. 3. AI can facilitate synthetic biology research and diagnostics development. Examples of AI model inputs, model architectures or types, and model outputs relevant to the development of synthetic biology–based diagnostics (**A**) and the development of other forms of diagnostics, including those based on sequencing, mass spectrometry, and imaging (**B**). Inputs, models, and outputs shown are representative, in part, of those discussed in (*55*–*60*, *68*–*74*).

models, which are trained to perform a specific task. Fine-tuned large language models for biology, such as BioBERT (*54*), may unify information from diverse infection contexts and offer increased predictive power to help elucidate host–pathogen interactions, facilitate antigen selection, inform vaccine design, and design treatment strategies.

AI for diagnostics and synthetic biology

As large-scale testing efforts during the COVID-19 pandemic have illustrated, quick and accurate detection of infections and pathogen outbreaks remains paramount to controlling the spread of infectious diseases. Recent advances in combining AI with synthetic biology, gene expression analyses, mass spectrometry, and imaging have substantively expanded our ability to detect infections and predict drug resistance (Fig. 3) (*55*–*60*). ML is well suited for catalyzing synthetic biology–based diagnostics because of the high programmability of biological elements, the routine generation of large or sequence-based datasets, and the ability of ML to extract meaningful information from biomolecular networks in disease biology (*61*).

Engineering genetic elements and understanding biomolecular networks remain critical to designs that harness biology. Synthetic biology approaches leveraging enzymatic reactions, toehold switches (RNAs that respond to specific nucleic acid sequences), or CRISPR-Cas enzymes have been used for the detection of malaria, Ebola, Zika, COVID-19, and other diseases (*62*–*67*). Supervised ML models have facilitated the design of toehold switches (*68*, *69*), CRISPR guide RNAs (*70*–*72*), and other biomolecules. Notably, large datasets are available for toehold switch function, CRISPR guide RNA activity, and other factors that are relevant to diagnostic design. Different types of neural networks, including feed-forward networks (neural networks with linear architectures), convolutional neural networks (networks composed of convolutional layers), and long short-term memory models, have been commonly used to model these data, but the same datasets can provide useful resources for testing more recently developed and potentially more predictive or generative ML models, including transformers or variational autoencoders (Box 1), to more efficiently develop next-generation diagnostics.

Beyond synthetic biology, ML has been used for gene expression–, mass spectrometry–, and imaging-based diagnostics. Gene expression– and mass spectrometry–based diagnostics have been applied to antimicrobial susceptibility testing (AST). AST remains important for informing the use of anti-infective drugs, but typical (culture-based) AST for bacteria, viruses, fungi, and parasites can take at least several days to complete. This turnaround time remains too long to adequately address clinical needs for acute systemic infections, such as those resulting in sepsis. Recent studies have combined gene expression and interaction profiling, structural mutation-mapping, and ML to identify genetic signatures of resistance that could be used as the basis of rapid molecular diagnostics (*56*, *57*). Supervised ML classifiers have predicted antibiotic resistance profiles correlated with clinical matrix-assisted laser desorption/ionization–time-of-flight (MALDI-TOF) mass spectra of bacterial proteins, and these predictions could be completed within 24 hours of sample collection (*58*). Nevertheless, a potential limitation to this approach is that the areas under the receiver operating characteristic curve (AUROC) for different bacterial species were ~0.7, suggesting that improvements in classifier accuracy will be needed to make this approach useful (e.g., AUROC > 0.9) in clinical settings. ML has also informed more-traditional ways of diagnosing infections, including microscopy, epitope profiling (*73*), chest radiographs and CT scans (*59*, *60*), and lateral flow tests (*74*). In each of these applications, the generation of large, multidimensional datasets combined with clear functional readouts, such as the presence or absence of a resistance profile or a disease, makes ML particularly useful for producing accurate predictions.

Nevertheless, there remain important challenges in applying ML to diagnosis, including low data quality or quantity for new or emerging pathogens, the limited generalizability of the current data and approaches used, and the need for highly accurate diagnostic predictions in clinical settings. Obtaining enough

high-quality data relevant to new or emerging pathogens or strains, particularly in low-resource settings, remains a difficult problem that is exacerbated by a lack of scientific infrastructure and variable public health resources. ML models based on limited data may exhibit biases, promulgating inappropriate diagnostics, misdiagnoses, and greater health inequalities that make it more difficult to serve patient populations. These biases may also remain undetected, especially when black box ML models, which do not provide any explanation or interpretation of their predictions, are used (*30*, *61*). Even when high-quality sequencing data from large infectious disease databases, such the PATRIC database (*75*), are available, it remains to be seen whether antimicrobial resistance predictions based on these data are generalizable when applied to genetically diverse infections found worldwide. Furthermore, unlike for anti-infective drug discovery—where the stakes for false positives and false negatives predicted by ML models are lower because the predictions can be further tested—the consequences of an inaccurate diagnostic prediction can be severe. In fact, a recent survey suggested that no existing model for the diagnosis or prognosis of COVID-19 from chest radiographs and CT scans was of potential clinical use owing to methodological flaws, biases, or both (*60*). Models with comparatively high AUROC values (i.e., 0.90) may still be too weak for clinical applications, as this value implies that, given a positive and a negative diagnosis, the negative diagnosis is ranked higher than the positive diagnosis 10% of the time. Until more-accurate ML models can be developed, AI-based diagnostics might play only a supporting role in clinical settings.

Moving forward, we anticipate that researchers will focus on the ML-guided design and discovery of synthetic circuits enabling the development of low-cost and portable diagnostics, the application of AI to data generated from clinical and field-deployable diagnostics that improve accessibility and scope, and the development of ML models that provide accurate diagnoses in clinical settings. In particular, the application of sequence-to-function models, language models, and generative models to RNA switches, CRISPR-based tools, and other programmable elements will be promising areas of growth given the ability for rapid iteration and the precise, on-target activity of these synthetic biology approaches (*67*–*71*). By increasing the testing and reporting of infections, the development of low-cost, field-deployable diagnostics should also help produce more-balanced datasets that better sample local infections and make ML models less biased. AI or ML models that can extract information from small or incomplete datasets, using tools such as transfer learning (which adapts models trained on a specific task to other tasks) and Bayesian networks (networks that allow for probabilistic inference), can play outsized roles in how infectious diseases are addressed, especially for overlooked populations in low-resource areas. Such models could lead to more-personalized medicine, in which diagnoses or resistance profiles can be readily reported on the basis of data from only a few infections and help guide the use of anti-infective drugs. On the other hand, the accuracy of ML models also needs to improve for practical use in clinical diagnoses. Future ML models will likely need to be optimized in architecture, thoroughly evaluated for biases, and trained on large amounts of robust data to achieve high accuracy. Transfer and multitask learning, attention mechanisms, and other approaches can help these next-generation ML models provide more-accurate diagnoses.

Outlook

Approaches combining systems and synthetic biology with ML models, including graph neural networks, sequence-to-function and sequence-to-structure frameworks, and generative models, are yielding access to drug candidates and methods for drug discovery. Supervised classifiers, unsupervised language models, and other ML models have produced biologically relevant insights into how pathogens interact with host cells and immune responses, informing antigen determination, vaccine design, and treatment strategies. The aforementioned types of ML models have also informed the design of various diagnostic tools and improved system accuracy, helping clinicians to diagnose infections and detect antimicrobial resistance. Beyond medical and biotechnological approaches to infectious diseases, ML—and AI more broadly—has also led to substantive advances in epidemiology and our understanding of disease transmission. Better leveraging of AI to address infectious diseases will require a collaborative effort among scientists, clinicians, and public health officials.

Developing AI models that generalize and avoid bias will require the acquisition and integration of comprehensive datasets. These datasets might include high-throughput therapeutic counter-screens and explorations of diverse chemical spaces for drug discovery, data from drug–target interactions and biomolecular interactions, and genetic sequencing information that is robustly and representatively sampled from all infections, including those occurring in low-resource or hard-to-access areas. Programmable modalities, such as nucleic acid and amino acid sequences, have represented tractable and common starting points for ML models (such as those predicting structure from sequence), but advances in biology and chemistry are important to opening up search spaces and making biology more "embeddable," or able to be represented by low-dimensional features. Progress in this area will help to predict therapeutic efficacy and drug MoAs, complex host–pathogen interactions and host responses, and interactions between small molecules, proteins, peptides, and nucleic acids. Advances in AI will include approaches, such as few-shot and multitask models, that leverage more of the available scientific information for dealing with limited or low-quality data. Furthermore, interpretable, explainable, and generative ML approaches will lead to specific biological hypotheses and insights. We anticipate that AI will continue to empower us to design next-generation drugs, vaccines, and diagnostics that address infectious diseases.

REFERENCES AND NOTES

1. T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
2. G. Schneider, *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
3. R. O'Shea, H. E. Moser, *J. Med. Chem.* **51**, 2871–2878 (2008).
4. J. G. Moffat, F. Vincent, J. A. Lee, J. Eder, M. Prunotto, *Nat. Rev. Drug Discov.* **16**, 531–543 (2017).
5. S. Afrasiabi, M. Pourhajibagher, R. Raoofian, M. Tabarzad, A. Bahador, *J. Biomed. Sci.* **27**, 6 (2020).
6. M. D. T. Torres *et al*., *Nat. Biomed. Eng.* **6**, 67–75 (2022).
7. M. D. T. Torres, J. Cao, O. L. Franco, T. K. Lu, C. de la Fuente-Nunez, *ACS Nano* **15**, 2143–2164 (2021).
8. M. Der Torossian Torres, C. de la Fuente-Nunez, *Chem. Commun.* **55**, 15020–15032 (2019).
9. J. M. Stokes *et al*., *Cell* **180**, 688–702.e13 (2020).
10. G. Liu *et al*., *Nat. Chem. Biol.* 10.1038/s41589-023-01349-8 (2023).
11. F. Wong, S. Omori, N. M. Donghia, E. J. Zheng, J. J. Collins, *Nat. Aging* **3**, 734–750 (2023).
12. M. C. R. Melo, J. R. M. A. Maasch, C. de la Fuente-Nunez, *Commun. Biol.* **4**, 1050 (2021).
13. F. Wan, D. Kontogiorgos-Heintz, C. de la Fuente-Nunez, *Digit. Discov.* **1**, 195–208 (2022).
14. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
15. M. Baek *et al*., *Science* **373**, 871–876 (2021).
16. P. D. Karp *et al*., *Brief. Bioinform.* **20**, 1085–1093 (2019).
17. P. D. Karp *et al*., *EcoSal Plus* **8**, 10.1128/ecosalplus.ESP-0006-2018 (2018).
18. K. L. Howe *et al*., *Nucleic Acids Res.* **48**, D689–D695 (2020).
19. C. Fu *et al*., *Nat. Commun.* **12**, 6497 (2021).
20. J. L. Espinoza *et al*., *PLOS Comput. Biol.* **17**, e1008857 (2021).
21. J. H. Yang *et al*., *Cell* **177**, 1649–1661.e9 (2019).
22. W. F. Porto *et al*., *Nat. Commun.* **9**, 1490 (2018).
23. J. R. M. A. Maasch, M. D. T. Torres, M. C. R. Melo, C. de la Fuente-Nunez, bioRxiv 2022.11.15.516443 [Preprint] (2022). https://doi.org/10.1101/2022.11.15.516443.
24. S. B. Jones, B. Simmons, A. Mastracchio, D. W. C. MacMillan, *Nature* **475**, 183–188 (2011).
25. L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
26. S. L. Rössler, N. M. Grob, S. L. Buchwald, B. L. Pentelute, *Science* **379**, 939–945 (2023).
27. A. J. Quartararo *et al*., *Nat. Commun.* **11**, 3183 (2020).
28. N. Iwano, T. Adachi, K. Aoki, Y. Nakamura, M. Hamada, *Nat. Comput. Sci.* **2**, 378–386 (2022).
29. H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **3**, 283–293 (2017).
30. J. Jiménez-Luna, F. Grisoni, G. Schneider, *Nat. Mach. Intell.* **2**, 573–584 (2020).
31. P. Das *et al*., *Nat. Biomed. Eng.* **5**, 613–623 (2021).
32. D. Nagarajan *et al*., *J. Biol. Chem.* **293**, 3492–3509 (2018).
33. W. Jin *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105070118 (2021).
34. F. Wong *et al*., *Nat. Commun.* **12**, 2321 (2021).
35. D. Lowe, "Why AlphaFold won't revolutionise drug discovery," *Chemistry World*, 5 August 2022; https://www.chemistryworld.com/opinion/why-alphafold-wont-revolutionise-drug-discovery/4016051.article.
36. F. Wong *et al*., *Mol. Syst. Biol.* **18**, e11081 (2022).
37. E. B. Breidenstein, C. de la Fuente-Núñez, R. E. Hancock, *Trends Microbiol.* **19**, 419–426 (2011).
38. A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori, M. S. Rao, *Chem. Res. Toxicol.* **33**, 20–37 (2020).
39. N. Palmer, J. R. M. A. Maasch, M. D. T. Torres, C. de la Fuente-Nunez, *Infect. Immun.* **89**, e00703-20 (2021).
40. A. M. Watkins, R. Rangan, R. Das, *Structure* **28**, 963–976.e6 (2020).

41. N. E. Wheeler, P. P. Gardner, L. Barquist, *PLOS Genet.* **14**, e1007333 (2018).
42. H. Chen *et al.*, *Brief. Bioinform.* **22**, bbaa068 (2021).
43. D. Bojar, R. K. Powers, D. M. Camacho, J. J. Collins, *Cell Host Microbe* **29**, 132–144.e3 (2021).
44. D. Fisch *et al.*, *eLife* **8**, e40560 (2019).
45. B. Hie, E. D. Zhong, B. Berger, B. Bryson, *Science* **371**, 284–288 (2021).
46. P. J. Sample *et al.*, *Nat. Biotechnol.* **37**, 803–809 (2019).
47. E. Ong *et al.*, *Bioinformatics* **36**, 3185–3191 (2020).
48. G. W. Ashdown *et al.*, *Sci. Adv.* **6**, eaba9338 (2020).
49. L. Boeck *et al.*, *Nat. Microbiol.* **7**, 1431–1441 (2022).
50. N. Chaudhary, D. Weissman, K. A. Whitehead, *Nat. Rev. Drug Discov.* **20**, 817–838 (2021).
51. M. C. Crank *et al.*, *Science* **365**, 505–509 (2019).
52. M. Stracy *et al.*, *Science* **375**, 889–894 (2022).
53. D. M. Cornforth *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5125–E5134 (2018).
54. J. Lee *et al.*, *Bioinformatics* **36**, 1234–1240 (2020).
55. H. C. Metsky *et al.*, *Nat. Biotechnol.* **40**, 1123–1131 (2022).
56. A. Khaledi *et al.*, *EMBO Mol. Med.* **12**, e10264 (2020).
57. E. S. Kavvas *et al.*, *Nat. Commun.* **9**, 4306 (2018).
58. C. Weis *et al.*, *Nat. Med.* **28**, 164–174 (2022).
59. X. Mei *et al.*, *Nat. Med.* **26**, 1224–1228 (2020).
60. M. Roberts *et al.*, *Nat. Mach. Intell.* **3**, 199–217 (2021).
61. D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, J. J. Collins, *Cell* **173**, 1581–1592 (2018).
62. L. F. de Lima, A. L. Ferreira, M. D. T. Torres, W. R. de Araujo, C. de la Fuente-Nunez, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2106724118 (2021).
63. K. Pardee *et al.*, *Cell* **159**, 940–954 (2014).
64. K. Pardee *et al.*, *Cell* **165**, 1255–1266 (2016).
65. M. Karlikow *et al.*, *Nat. Biomed. Eng.* **6**, 246–256 (2022).
66. R. A. Lee *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25722–25731 (2020).
67. H. de Puig *et al.*, *Sci. Adv.* **7**, eabh2944 (2021).
68. N. M. Angenent-Mari, A. S. Garruss, L. R. Soenksen, G. Church, J. J. Collins, *Nat. Commun.* **11**, 5057 (2020).
69. J. A. Valeri *et al.*, *Nat. Commun.* **11**, 5058 (2020).
70. G. Chuai *et al.*, *Genome Biol.* **19**, 80 (2018).
71. H. K. Kim *et al.*, *Nat. Biotechnol.* **36**, 239–241 (2018).
72. D. Wang *et al.*, *Nat. Commun.* **10**, 4284 (2019).
73. E. Shrock *et al.*, *Science* **370**, abd4250 (2020).
74. V. Turbé *et al.*, *Nat. Med.* **27**, 1165–1170 (2021).
75. A. R. Wattam *et al.*, *Nucleic Acids Res.* **42**, D581–D591 (2014).

ACKNOWLEDGMENTS

We thank X. Tan, M. N. Anahtar, and J. A. Valeri for helpful comments on the manuscript. **Funding:** F.W. was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number K25AI168451. C.F.N. holds a Presidential Professorship at the University of Pennsylvania, is a recipient of the Langer Prize awarded by the AIChE Foundation, and acknowledges funding from the IADR Innovation in Oral Care Award, the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, a Penn Health-Tech Accelerator Award, the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania, the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138201, and the Defense Threat Reduction Agency (HDTRA11810041, HDTRA1-21-1-0014, and HDTRA1-23-1-0001). J.J.C. was supported by the Defense Threat Reduction Agency (HDTRA12210032), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01-AI146194, and the Broad Institute of MIT and Harvard. This work is part of the Antibiotics-AI Project, which is directed by J.J.C. and supported by the Audacious Project; Flu Lab, LLC; the Sea Grape Foundation; and Rosamund Zander and Hansjorg Wyss for the Wyss Foundation. **Competing interests:** J.J.C. is scientific cofounder and scientific advisory board chair of EnBiotix, an antibiotic drug discovery company, and Phare Bio, a nonprofit venture focused on antibiotic drug development. C.F.N. provides consulting services to Invaio Sciences and is a member of the scientific advisory boards of Nowture S.L. and Phare Bio. F.W. declares no competing interests.

Submitted 1 April 2023; accepted 5 June 2023
10.1126/science.adh1114

Leveraging artificial intelligence in the fight against infectious diseases

Felix Wong, Cesar de la Fuente-Nunez, and James J. Collins

Science, **381** (6654), .
DOI: 10.1126/science.adh1114