Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

A brief guide to machine learning for antibiotic discovery

Gary Liu[1,2,3] and Jonathan M Stokes[1,2,3]

Rising antibiotic resistance and an alarmingly lean antibiotic pipeline require the adoption of novel approaches to rapidly discover new structural and functional classes of antibiotics. Excitingly, algorithmic approaches to antibiotic discovery are sufficiently advanced to meaningfully influence the antibiotic discovery process. Indeed, once trained on high-quality datasets, contemporary machine-learning and deep-learning models can be used to perform predictions for new antibiotics across vast chemical spaces, orders of magnitude more rapidly than compounds can be screened in the laboratory. This increases the probability of discovering new antibiotics with desirable properties. In this short review, we briefly describe the utility of contemporary machine-learning and deep-learning approaches to guide the discovery of new small-molecule antibiotics and unidentified natural products. We then propose a call to action for more open sharing of high-quality screening datasets to accelerate the rate at which forthcoming antibiotic-prediction models can be trained. Together, we aim to introduce antibiotic discoverers to a sample of recent applications of contemporary algorithmic methods to facilitate the wider adoption of these powerful computational approaches.

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

Corresponding author: Jonathan M Stokes (stokesjm@mcmaster.ca)

Current Opinion in Microbiology 2022, **69**:102190

This review comes from a themed issue on **Antimicrobials**

Edited by **David Rogers** and **Richard E Lee**

https://doi.org/10.1016/j.mib.2022.102190

Introduction

Since the discovery of penicillin, antibiotics have become the cornerstone of modern medicine. Unfortunately, the continued efficacy of these essential drugs is at severe risk due to the persistent global dissemination of antibiotic-resistance determinants. Moreover, the decreasing development of novel antibiotics in the private sector is exacerbating this already- dire problem. Indeed, in 2019, there were an estimated 4.95 million deaths associated with antibiotic resistance [1••]. Without immediate action to discover and develop new antibiotics, it is projected that deaths from resistant infections will reach 10 million per year by 2050, surpassing even cancer [2].

Most of our clinically used antibiotics were discovered between the 1940s and 1960s by screening secondary metabolites produced by soil-dwelling microbes for those that displayed antibacterial activity *in vitro* [3,4]. Unfortunately, this systematic screening approach — the Waksman platform — experienced a shortcoming in the form of the dereplication problem, wherein investigators were repeatedly discovering the same antibiotics [5]. Therefore, in more recent decades, antibiotic discoverers have turned to medicinal chemistry approaches [6] to modify existing antibiotic scaffolds, as well as high-throughput chemical screening to discover new ones [7]. Medicinal chemistry has afforded us the ability to modify existing antibiotics to optimize medicinal and antibacterial properties in the face of rising resistance [8]. Indeed, this is the primary method that has led to 'new' antibiotics over the past few decades. However, such approaches are not ideally suited to discover fundamentally novel antibiotic scaffolds.

Conversely, high-throughput chemical screening has proven useful in identifying an array of fundamentally novel antibacterial small molecules with activity *in vitro*, but has failed to result in any new antibiotics suitable for clinical application. It is likely a combination of two factors that contributed to the failure of high-throughput screening to result in new clinical antibiotics [9]. The first is the number of chemicals that can be empirically screened in the laboratory, which is upper bounded by a few million molecules. While at first consideration this seems like a reasonable number of compounds for primary screening, when compared with the theoretical number of compounds that display drug-like properties ($\sim 10^{60}$), these screens are narrow [10]. The second factor is the somewhat unsuitable chemical space that is often explored in synthetic screening libraries. Synthetic compound libraries are generally built using combinatorial chemistry methods centered around molecular properties that make for human-targeting drugs, but not

antibiotics, which commonly occupy distinct chemical spaces [11,12].

Excitingly, algorithmic methods to antibiotic discovery are reaching a state of performance that can meaningfully influence how novel antibacterial molecules are discovered [13•,14]. *In silico* approaches afford users the ability to explore vast chemical spaces — upward of tens of billions of molecules or more — much more rapidly than is possible using wet-lab experimentation. This increases the probability of identifying structurally and functionally novel chemicals, with molecular properties that are amenable for further development into antibiotic drugs. However, contemporary machine-learning techniques require large quantities of diverse and high-quality data with which to learn [15]. It is therefore essential that investigators aiming to leverage machine-learning methods understand the importance of appropriate training data acquisition, as well as the optimal machine-learning model architecture for a given prediction task. With some simple guidelines in mind, we posit that machine-learning approaches to antibiotic discovery — and drug discovery in general — can become widely adopted, accelerating the rate at which novel antibiotics are discovered and advanced into the clinic.

In this short review, we highlight some modern examples of the application of machine learning toward antibiotic discovery. We will first describe contemporary machine-learning approaches to aid in the discovery of novel small-molecule antibiotics. This will lead into a brief discussion of algorithmic approaches for natural product discovery. Last, we will close the paper by proposing a call to action for open data sharing, allowing for wider access to well-defined and high-quality training data for everyone in the antibiotic discovery field.

Machine learning for small-molecule antibiotic discovery

The notion of leveraging computed chemical features to guide molecular property prediction dates to the development of quantitative structure–activity relationships (QSAR) in the mid-20th century [16]. This mathematical framework provided an opportunity to quantitatively associate computable molecular features with physicochemical properties or biological activities. The classical QSAR approach applies expert knowledge to build useful representations — so-called molecular fingerprints — of the graph structures of molecules [17]. Such features defined in fingerprints may include molecular weight, hydrophobicity, number of rotatable bonds, polar surface area, and the list can continue. These expert-defined molecular representations can then be used as inputs to train classification or regression models to predict physicochemical or biological properties of interest.

As a concrete example, Wang et al. [18] described the application of various machine-learning models — naive Bayes (described in [19]), support vector machine (described in [20,21]), recursive-partitioning (described in [22]), and k-nearest neighbor (described in [23]) — trained on expert-defined molecular descriptors to predict new antibiotics with activity against *Staphylococcus aureus*. Specifically, the authors applied a collection of ~30 computable molecular descriptors to ~5000 molecules and used these as inputs to train their collection of machine-learning models. They subsequently applied their highest-performing model to a collection of ~7500 naive compounds and validated 12 molecules as antibacterial against *S. aureus in vitro*. Similarly, Li et al. [24] employed naive Bayes and recursive-partitioning machine-learning models to predict antibacterial molecules targeting DNA gyrase in *Escherichia coli* and *S. aureus*. Here, the authors applied 51 molecular descriptors to build fingerprints of a modest 137-compound training dataset, then applied the best-performing model to an in-house collection of 488 compounds. This resulted in the acquisition of four structurally similar compounds that validated as antibacterial *in vitro*.

We note here that while it is likely that more contemporary deep-learning techniques [25] would have outperformed these simpler methods, particularly in the context of generalization into new chemical spaces, these studies nevertheless show that it is possible to identify novel antibacterial molecules using easily computable fingerprints and simple classifiers. However, conventional machine-learning algorithms trained on computable fingerprint vectors are intrinsically limited to a relatively narrow spectrum of human-defined molecular features. This is in stark contrast to contemporary deep-learning architectures that learn the optimal molecular representations for a specific prediction task automatically [26]. Indeed, such deep-learning approaches have been shown to outperform traditional machine-learning-based approaches on an array of physicochemical and biological property prediction tasks and can more robustly generalize to chemical spaces beyond that on which the model was originally trained [13•,27••].

As a concrete example, a recent application of contemporary deep learning for novel antibiotic discovery was described by Stokes et al. [27••], where they leveraged a deep-learning architecture called a message-passing neural network (MPNN, described in [13•]) to discover structurally novel antibiotics against laboratory strain *E. coli*. Here, the authors trained their MPNN on a collection of ~2500 small molecules for those that inhibited the growth of *E. coli in vitro*. This screening data was then binarized into compounds that were growth-inhibitory (labeled as 1) and those that were not (labeled as 0). The MPNN took as input the graph structures of the chemicals — including simple features [28] of the

atoms and bonds within each molecule — from this training dataset, along with their corresponding activity values (0 or 1), then automatically learned how to convert the structures of chemicals into vector representations that maximized the ability of the model to correctly classify molecules as growth-inhibitory or not. Importantly, in contrast to previous methods, no expert-defined molecular features were strictly required to build the molecular representations. Next, this trained model was shown a collection of ~107 million structurally diverse chemicals from various *in silico* chemical repositories [29•,30] — most of which resided in distinct chemical spaces relative to the training dataset — and outputted a prediction value for each, reflecting the model's interpretation of whether a given molecule was antibacterial. This experimental and computational pipeline resulted in the discovery of numerous unique molecules with activity against *E. coli* and other phylogenetically diverse pathogenic bacterial species.

Encouraging the broad adoption of algorithmic approaches for small-molecule antibiotic discovery — whether simple machine-learning models trained using conventional molecular fingerprint vectors or more sophisticated deep-learning models that learn molecular representations automatically (Figure 1a) — we emphasize here the importance of two aspects of the combined experimental/computational pipeline. First is the curation of a chemically diverse training dataset generated using highly controlled experimental conditions. The goal during training is to provide the model with many structurally diverse chemicals with varying levels of antibacterial activity. This will allow for maximum generalization when performing predictions on *in silico* chemical sets. Here, it is important that these training data be acquired using well-defined experimental conditions to ensure that wet-lab validation of *in silico* molecular predictions is performed using the experimental parameters on which the model was trained. This is the most informative approach to understand 'real world' model performance. Indeed, investigators must use great caution when training molecular property prediction models using publicly available primary screening datasets, since the methodologies for acquisition of these data can be insufficiently defined. Second is the use of appropriate *in silico* filtering rules to prioritize predicted chemicals for wet-lab validation. For example, when performing predictions on hundreds of millions to billions of molecules, it is common to retrieve tens of thousands of molecules that pass a given prediction-score threshold. To reduce this number to hundreds or thousands of molecules for wet-lab validation, investigators must develop additional criteria to acquire top candidates [31]. Such secondary filters may include structural divergence from training set chemicals, structural diversity among the prioritized compounds themselves, ease of synthetic accessibility, or predicted nontoxicity to human cells. In any case, investigators must be careful to avoid excluding valuable chemicals through the application of overly restrictive inclusion criteria.

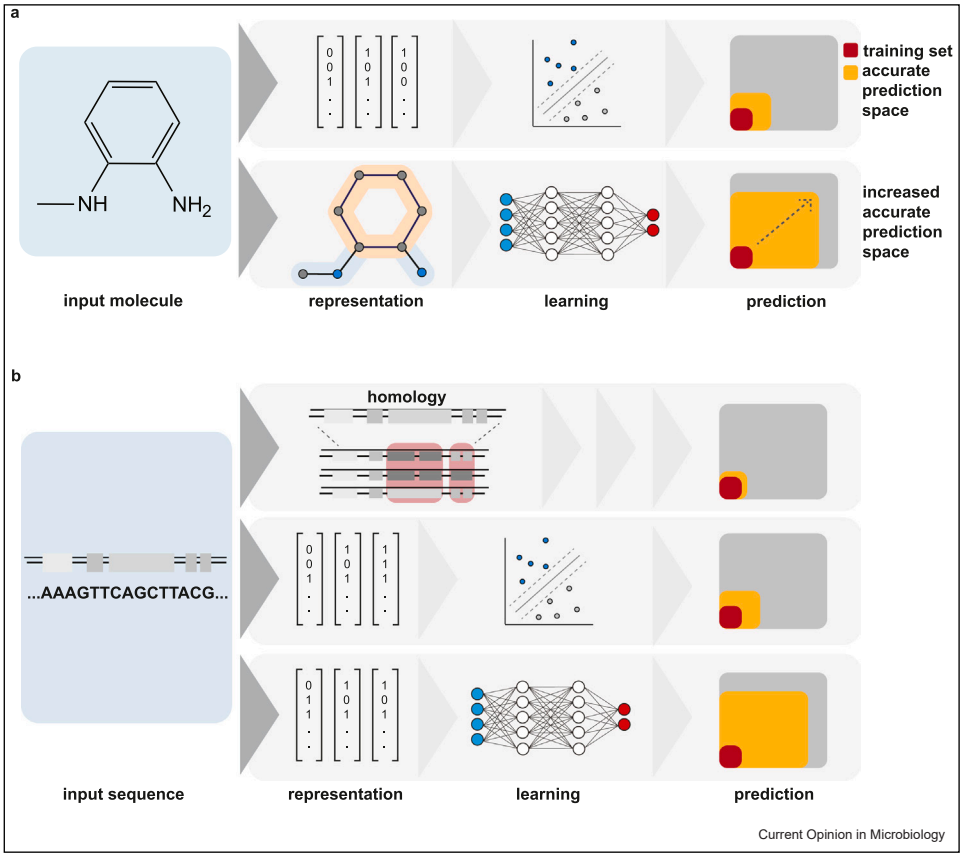
Machine learning for natural product antibiotic discovery

Most of our current clinical antibiotics are derived from nature [32]. While the Waksman era of antibiotic discovery has ended — or at least paused — the probability that we have identified all clinically useful antibacterial natural products is low. It is far more likely that unidentified natural product antibiotics with clinical utility are synthesized by so-called cryptic biosynthetic gene clusters (BGC) — genetic elements that are repressed in standard laboratory conditions [33,34]. Recent work at the intersection of natural product discovery and machine learning has aimed to predict BGCs from genomics data, as well as the bioactivities of BGC-encoded compounds (Figure 1b). Excitingly, the exponential growth of whole-genome sequencing data has catalyzed the development of computational approaches with which to predict the medicinal importance of sequences therein.

Classical predictive tools rely on sequence homology between putative BGCs and characterized BGCs [35–37]. While this approach is adequate to identify functionally related BGCs, these methods can fail to generalize to new sequences — and therefore new natural products — that may display antibacterial activity. To overcome this inherent limitation of homology-based models, machine-learning and deep-learning models have been developed to improve generalizability toward the discovery of novel classes of BGCs [38,39]. For example, Hannigan et al. [40•] leveraged a recurrent neural network-based architecture to develop DeepBGC, which can predict biosynthetic gene clusters from bacterial genomes. Indeed, after training DeepBGC on ~600 positive BGC examples and thousands of negative BGC examples, their prediction pipeline could identify BGC candidates for molecules with putative antibiotic activity that were unidentified by basic rule-based methods. Moreover, their data suggest that DeepBGC displayed a greater potential to generalize and identify gene-cluster classes that it had not directly observed during training.

Progress has also been made to predict the antibacterial activities of natural products based on BGC sequence using machine-learning methods [41,42]. For instance, Walker and Clardy [43] developed a binary classifier that takes as input BGC sequence data, converts each BGC into an expert-defined vector based on the presence/absence of BGC elements, and outputs whether a given BGC is associated with a natural product bioactivity of interest using common machine-learning models (random forest, support vector machine, and logistic regression). While their classifier performed well on BGCs

Figure 1



The utility of algorithmic approaches to novel antibiotic discovery. **(a)** Classical machine-learning-based approaches to small-molecule antibiotic discovery rely on expert-defined molecular property extraction and fail to generalize well to chemical spaces far beyond that of the initial training dataset (top). However, contemporary deep-learning approaches that generate molecular representations automatically have been shown to generalize well to novel chemical spaces, increasing the probability of discovering structurally and functionally novel antibiotics (bottom). **(b)** Classical methods to identify BGCs relied on sequence homology to known BGCs of interest and therefore limited the scope of sequence space that could be explored for novel BGC and natural product discovery (top). More recent representation-based methods using machine-learning (middle), or deep-learning (bottom) methods can generalize to new sequence spaces more robustly, enhancing their ability to discover novel BGCs of potential interest for new natural product antibiotics.

for which many examples were observed during training (~80% maximum accuracy), like most machine-learning-based approaches, performance suffered when more generalization was required (~60% minimum accuracy). Future deep-learning model development, which does not rely on expert-defined features for BGC encoding, will improve generalization into new BGC sequence space, allow for more robust structural predictions based on BGC sequence, and enable more accurate predictions of the antibacterial efficacy of these natural products based on property- prediction models such as those described in the previous section.

Call to action

We posit that the antibiotic discovery field is ripe for the widespread adoption of algorithmic solutions for antibiotic discovery — whether small molecules and natural products, or antibiotic alternatives that are outside the scope of this short review, such as antimicrobial peptides [44,45] and phage [46,47]. However, contemporary machine-learning and deep-learning methods do not exist in isolation beyond the wet lab. Contrarily, these algorithmic methods require large amounts of high-quality data on which to train. Currently, insufficient emphasis is placed on publishing the entirety of screening datasets with antibiotic discovery studies. The lack of well-described, easily accessible, and open-source datasets is hindering the rate at which machine-learning and deep-learning tools are built and employed for antibiotic-prediction tasks. Therefore, we urge members of the antibiotic discovery community in academia and industry to engage in open dissemination of their antibiotic screening datasets to allow for the collective training of robust antibiotic-prediction models that can

maximally generalize to new chemical spaces. This will accelerate the rate at which we are able to discover and optimize novel structural and functional classes of antibiotics from vast *in silico* chemical and sequence spaces.

Conflict of interest statement

J.M.S. is co-founder and scientific director of Phare Bio.

Acknowledgements

This paper was supported by the David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada, Weston Family Foundation, Toronto, Ontario, Canada. We thank Ryan Tso for artistic input for Figure 1.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest.

1. •• Murray CJL, *et al.*: **Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis**. *Lancet* 2022, **399**:629-655.

This paper provides a quantitative assessment of the severity of the modern antibiotic resistance crisis.

2. Review on Antimicrobial Resistance. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations; 2014 https://amr-review.org/sites/default/files/AMR%20Review%20Paper%20-%20Tackling%20a%20crisis%20for%20the%20health%20and%20wealth%20of%20nations_1.pdf.

3. Brown ED, Wright GD: **Antibacterial drug discovery in the resistance era**. *Nature* 2016, **529**:336-343.

4. Lewis K: **Antibiotics: recover the lost art of drug discovery**. *Nature* 2012, **485**:439-440.

5. Cox G, *et al.*: **A common platform for antibiotic dereplication and adjuvant discovery**. *Cell Chem Biol* 2017, **24**:98-109.

6. Pawlowski AC, Johnson JW, Wright GD: **Evolving medicinal chemistry strategies in antibiotic discovery**. *Curr Opin Biotechnol* 2016, **42**:108-117.

7. Brown DG, May-Dracka TL, Gagnon MM, Tommasi R: **Trends and exceptions of physical properties on antibacterial activity for Gram-positive and Gram-negative pathogens**. *J Med Chem* 2014, **57**:10144-10161.

8. Velkov T, Thompson PE, Nation RL, Li J: **Structure-activity relationships of polymyxin antibiotics**. *J Med Chem* 2010, **53**:1898-1916.

9. Tommasi R, Brown DG, Walkup GK, Manchester JI, Miller AA: **ESKAPEing the labyrinth of antibacterial discovery**. *Nat Rev Drug Discov* 2015, **14**:529-542.

10. Reymond J-L: **The chemical space project**. *Acc Chem Res* 2015, **48**:722-730.

11. Liu R, Li X, Lam KS: **Combinatorial chemistry in drug discovery**. *Curr Opin Chem Biol* 2017, **38**:117-126.

12. Blaskovich MAT, Zuegg J, Elliott AG, Cooper MA: **Helping chemists discover new antibiotics**. *ACS Infect Dis* 2015, **1**:285-287.

13. • Yang K, *et al.*: **Analyzing learned molecular representations for property prediction**. *J Chem Inf Model* 2019, **59**:3370-3388.

The authors describe the development of a powerful message passing deep neural network architecture for molecular property prediction tasks.

14. Walters WP, Barzilay R: **Critical assessment of AI in drug discovery**. *Expert Opin Drug Discov* 2021, **16**:937-947.

15. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K: **Efficient machine learning for big data: a review**. *Big Data Res* 2015, **2**:87-93.

16. Hansch C, Maloney PP, Fujita T, Muir RM: **Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients**. *Nature* 1962, **194**:178-180.

17. Cherkasov A, *et al.*: **QSAR modeling: where have you been? Where are you going to?** *J Med Chem* 2014, **57**:4977-5010.

18. Wang L, *et al.*: **Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches**. *J Chem Inf Model* 2014, **54**:3186-3197.

19. Zhang H, *et al.*: **Development of novel in silico prediction model for drug-induced ototoxicity by using naïve Bayes classifier approach**. *Toxicol Vitro* 2020, **65**:104812.

20. Li H, Liang Y, Xu Q: **Support vector machines and its applications in chemistry**. *Chemom Intellig Lab Syst* 2009, **95**:188-198.

21. Heikamp K, Bajorath J: **Support vector machines for drug discovery**. *Expert Opin Drug Discov* 2014, **9**:93-104.

22. Lamana C, Bellini M, Padova A, Westerberg G, Maccari L: **Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process**. *J Med Chem* 2008, **51**:2891-2897, https://doi.org/10.1021/jm701407x

23. Nigsch F, *et al.*: **Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization**. *J Chem Inf Model* 2006, **46**:2412-2422.

24. Li L, *et al.*: **Discovering new DNA gyrase inhibitors using machine learning approaches**. *RSC Adv* 2015, **5**:105600-105608.

25. Wang Z, *et al.*: **Advanced graph and sequence neural networks for molecular property prediction and drug discovery**. *Bioinformatics* 2022, **38**:btac112.

26. Walters WP, Barzilay R: **Applications of deep learning in molecule generation and molecular property prediction**. *Acc Chem Res* 2021, **54**:263-270.

27. •• Stokes JM, *et al.*: **A deep learning approach to antibiotic discovery**. *Cell* 2020, **180**:688-702 e13.

This study applied a message passing neural network trained on a purpose-built training datset to predict structurally novel antibacterial molecules against *E. coli*.

28. Landrum, G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling.https://www.rdkit.org/RDKit_Overview.pdf.

29. • Sterling T, Irwin JJ: **ZINC 15-ligand discovery for everyone**. *J Chem Inf Model* 2015, **55**:2324-2337.

This paper describes a searchable virtual repository of nearly 1.5 billion molecules that can be purchased or synthesized; it is a valuable resource for investigators interested in leveraging algorithmic approaches for molecular property prediction tasks.

30. Corsello SM, *et al.*: **The Drug Repurposing Hub: a next-generation drug library and information resource**. *Nat Med* 2017, **23**:405-408.

31. Hughes JP, Rees S, Kalindjian SB, Philpott KL: **Principles of early drug discovery**. *Br J Pharmacol* 2011, **162**:1239-1249.

32. Walsh C: **Where will new antibiotics come from?** *Nat Rev Microbiol* 2003, **1**:65-70.

33. Gupta A, *et al.*: **Global awakening of cryptic biosynthetic gene clusters in Burkholderia thailandensis**. *ACS Chem Biol* 2017, **12**:3012-3021.

34. Scherlach K, Hertweck C: **Mining and unearthing hidden biosynthetic potential**. *Nat Commun* 2021, **12**:3864.

35. Ren H, Shi C, Zhao H: **Computational tools for discovering and engineering natural product biosynthetic pathways**. *iScience* 2020, **23**:100795.

36. Hooft JJJ van der, *et al*.: **Linking genomics and metabolomics to chart specialized metabolic diversity**. *Chem Soc Rev* 2020, **49**:3297-3314.

37. Montalbán-López M, *et al*.: **New developments in RiPP discovery, enzymology and engineering**. *Nat Prod Rep* 2021, **38**:130-239.

38. Cimermancic P, *et al*.: **Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters**. *Cell* 2014, **158**:412-421.

39. Kloosterman AM, *et al*.: **Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides**. *PLoS Biol* 2020, **18**:e3001026.

40. • Hannigan GD, *et al*.: **A deep learning genome-mining strategy for biosynthetic gene cluster prediction**. *Nucleic Acids Res* 2019, **47**:e110 https://academic.oup.com/nar/article/47/18/e110/5545735.

This paper presents a deep learning pipeline to discover novel biosynthetic gene clusters with improved ability to generalize to new sequence space relative to conventional machine learning methods.

41. Prihoda D, *et al*.: **The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability**. *Nat. Prod Rep* 2021, **38**:1100-1108.

42. Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F: **Natural product drug discovery in the artificial intelligence era**. *Chem Sci* 2022, **13**:1526-1546.

43. Walker AS, Clardy J: **A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters**. *J Chem Inf Model* 2021, **61**:2560-2571.

44. Torres MDT, Melo MCR, Crescenzi O, Notomista E, de la Fuente-Nunez C: **Mining for encrypted peptide antibiotics in the human proteome**. *Nat Biomed Eng* 2022, **6**:67-75.

45. Lei J, *et al*.: **The antimicrobial peptides and their potential clinical applications**. *Am J Transl Res* 2019, **11**:3919-3931.

46. Kortright KE, Chan BK, Koff JL, Turner PE: **Phage therapy: a renewed approach to combat antibiotic-resistant bacteria**. *Cell Host Microbe* 2019, **25**:219-232.

47. Gu Liu C, *et al*., Green SI, Min L, Clark JR, Salazar KC, Terwilliger AL, Kaplan HB, Trautner BW, Ramig RF, Maresso AW: **Phage-antibiotic synergy is driven by a unique combination of antibacterial mechanism of action and stoichiometry**. *MBio* 2020, **11**:e01462-20, , https://doi.org/10.1128/mBio.01462-20