



Journal of Advanced Pharmaceutical Sciences and Natural Products (JAPSNP)

NATURAL PRODUCT-INSPIRED DRUG DISCOVERY REVISITED: GENOME MINING, SYNTHETIC BIOLOGY AND ARTIFICIAL INTELLIGENCE FOR REVIVING AN UNDEREXPLOITED CHEMICAL SPACE

Dr. Megha Verma*¹

Anupam Jaiswal¹

Mahima Beohar^{1,2}

¹*Department of Pharmacy, Gyan Ganga Institute of Technology & Sciences, Tilwara Ghat, Jabalpur, Madhya Pradesh 482003, India*

²*Department of Pharmaceutical Sciences, Shalom Institute of Health and Allied Sciences, Sam Higginbottom University of Agriculture, Technology & Sciences (SHUATS), Naini, Prayagraj, Uttar Pradesh 211007, India*

Corresponding author:

Dr. Megha Verma*

Email: anupamjaiswal@ggits.org

ABSTRACT

Natural products built most of the pharmacopoeia, yet the discovery engine that produced penicillin, the statins and the majority of clinically used antibiotics largely stalled by the late 1990s. The reason was not that microbes had run out of chemistry. Bioactivity-guided fractionation kept returning the same abundant molecules, and the enormous biosynthetic capacity of uncultured organisms stayed out of reach. A compendium of roughly 170,000 bacterial genomes estimates that only about three percent of that biosynthetic potential has ever been characterised.

This review argues that the field is being rebuilt around that gap by three technologies that matured separately and are now converging: genome mining, which reads biosynthetic gene clusters directly from sequence; synthetic biology, which expresses those clusters in tractable hosts; and machine learning, which prioritises both the clusters worth expressing and the molecules worth testing. Each is examined for what it has actually delivered, from antiSMASH-driven cluster detection and evolution-guided antibiotic discovery to the heterologous production of artemisinic acid and the deep-learning identification of halicin and a new structural class active against methicillin-resistant *Staphylococcus aureus*. The central claim is that the rate-limiting step has shifted from finding biosynthetic novelty to expressing and prioritising it, and that near-term returns depend on integrating the three approaches rather than advancing any one. The appraisal is deliberately uneven, because the evidence is: genome mining is mature, heterologous expression remains the chronic bottleneck, and artificial intelligence is powerful for property prediction but constrained by scarce, biased training data and a wide gap between an in silico hit and a validated drug.

Keywords: biosynthetic gene clusters; genome mining; synthetic biology; machine learning; antibiotic discovery; secondary metabolites.



1. Introduction

The case for natural products is quantitative, not sentimental. Of all small-molecule drugs approved between 1981 and 2019, close to two-thirds are natural products, direct derivatives, or synthetic molecules built on a natural-product scaffold or pharmacophore ¹. For anti-infectives and anticancer agents the dependence is heavier still. Natural products occupy regions of chemical space that combinatorial libraries rarely reach: more stereocentres, more oxygen, more conformational rigidity, and structures already selected over evolutionary time to engage biological targets ². No synthetic design programme has reproduced that hit rate against novel targets.

So the retreat of the pharmaceutical industry from natural-product screening around the turn of the century was never a verdict on the chemistry. It was a verdict on the workflow. Classical discovery grew a microbe, extracted it, fractionated the extract, and chased each activity to a pure compound. The approach was slow, it depended on culturing the producer, and above all it kept rediscovering the same high-abundance metabolites. Dereplication, the work of recognising a known compound before wasting months re-isolating it, became the dominant cost. As the obvious molecules were exhausted, the probability that any new extract held something genuinely new fell toward zero, and screening collections of synthetic compounds looked cheaper per assay. The decision to walk away was rational on the economics of the method available at the time.

Two findings reframed the problem. First, the molecules a microbe makes in a flask are a small fraction of the molecules its genome encodes. A global analysis of prokaryotic biosynthetic gene clusters, the physically clustered sets of genes that each assemble one metabolite, found biosynthetic diversity far exceeding the catalogue of known products, with individual actinomycete genomes carrying twenty to forty clusters where classical work had isolated only a handful ³. Most clusters are silent, or expressed too weakly to detect under laboratory conditions. Second, the producers themselves are mostly uncultured. The compendium of specialised-metabolite diversity assembled from around 170,000 bacterial genomes concluded that roughly 97 percent of the gene cluster families it could define had no known product, leaving only about three percent of the biosynthetic potential characterised ⁴. The chemistry was never exhausted. It was hidden, either inside silent clusters or inside organisms that do not grow on a plate.

This review takes the position that three technologies are now dismantling that bottleneck in sequence. Genome mining locates and classifies biosynthetic potential from DNA alone, so discovery no longer waits on expression in the native host. Synthetic biology moves a chosen cluster into a host built to express it, so a silent or uncultured pathway can be switched on deliberately. Machine learning sits across both, predicting which clusters encode novel or bioactive scaffolds and which finished molecules are worth synthesising and assaying. The



sections that follow treat each in turn, judge what it has delivered against what has been claimed for it, and then ask what their convergence actually buys. The argument is that the field's limiting step has moved downstream, from discovery to expression and prioritisation, and that the honest story of artificial intelligence here is one of genuine acceleration bounded by hard limits on data.

2. Genome mining: reading the biosynthetic blueprint

Genome mining rests on a structural fact of microbial biology. The genes that build a complex secondary metabolite usually sit together on the chromosome, and the enzymes they encode, modular polyketide synthases, nonribosomal peptide synthetases, terpene cyclases and their tailoring enzymes, follow recognisable sequence logic. A polyketide synthase module that loads and condenses one extender unit looks, in sequence, much like every other such module. That regularity is what makes clusters findable by algorithm, and it is why a genome sequence can be read as a parts list of the chemistry an organism is capable of making whether or not it ever makes it.

antiSMASH turned this into routine practice. The platform scans an assembled genome with curated detection rules and profile hidden Markov models, marks the boundaries of each cluster, assigns it to one of more than seventy classes, and for the better-understood assembly lines predicts a partial chemical structure of the product⁵. Its seventh release extended detection,

added regulatory annotation, and strengthened comparison against clusters of known function, and the platform now underpins a large fraction of published discovery work⁵. The methodological reviews catalogue the complementary tools that grew up around it, from precursor-peptide finders for ribosomally synthesised and post-translationally modified peptides to self-resistance locators⁶. Ribosomally synthesised peptides deserve particular mention: because they are gene-encoded rather than assembled by giant enzymes, their products can often be read more directly from sequence, and they have become one of the most productive classes for genomic discovery. Genome mining is, by any reasonable measure, the mature member of this trio.

Detection alone does not tell a chemist whether a cluster is worth pursuing, because most hits resemble something already known. Two developments address that. The first is reference infrastructure. MIBiG curates experimentally characterised clusters with their products and, where known, their biosynthetic logic, giving every new hit a benchmark to be compared against⁷. BiG-SCAPE then organises thousands of clusters into sequence-similarity networks and gene cluster families, so a mining campaign can see at a glance which hits fall into densely populated, well-mined families and which sit alone in sparse corners of biosynthetic space⁸. Rarity in that network is the signal. A cluster with no close characterised relative is the one most likely to encode a genuinely new scaffold,



and network position converts a flat list of hundreds of hits into a ranked one.

The second development is to let a hypothesis about function guide the search rather than casting a wide net. Resistance-gene-directed mining looks for clusters that carry a duplicated, modified copy of the very gene the product would inhibit, on the reasoning that a producer must protect itself from its own weapon, so a self-resistance gene points at both a novel compound and its molecular target. Evolution-guided mining applies the same logic phylogenetically. Culp and colleagues searched for glycopeptide clusters that had lost the genes responsible for the classical vancomycin mechanism, reasoning that any surviving, actively maintained cluster must act some other way, and recovered corbomycin and complestatin, antibiotics that kill by blocking the action of autolysins on the cell wall rather than by sequestering its building blocks⁹. That result matters beyond the two molecules. It shows genome mining used not as a net but as a hypothesis-driven instrument, targeting novel mechanism rather than novel structure for its own sake, which is the harder and more valuable target.

The scale on which this now operates is worth pausing on, because it changes the character of the problem. Public sequence databases hold hundreds of thousands of bacterial genomes and a rising volume of metagenome-assembled genomes recovered directly from soil, marine and host-associated communities without any culturing step. Mining those metagenomes is

how the uncultured majority becomes visible at all, and it is where the sparse, unfamiliar cluster families cluster. It also inverts the old constraint. Where classical discovery was starved of material, modern mining is flooded with candidates, and the entire downstream enterprise exists to decide which of them to act on. Abundance, not scarcity, is the governing condition.

The limits are equally clear, and worth stating rather than softening. A predicted cluster is a prediction, not a molecule. Structure prediction is reliable only for the most regular assembly lines, boundary calls between adjacent clusters are imperfect, and the single largest problem is that detecting a cluster says nothing about whether it is ever transcribed. That is the handoff to synthetic biology. It is also where the first machine-learning methods entered, with tools such as DeepBGC using recurrent neural networks over protein-family domains to recover clusters that rule-based detection misses and to lower the false-positive rate on unusual architectures¹⁰. DeepBGC is discussed with the other learning methods below, but its origin inside the mining literature is itself the point: the boundary between genome mining and artificial intelligence had already dissolved before anyone drew it.

3. Synthetic biology: from blueprint to molecule

If a genome encodes a promising cluster that stays silent, the problem becomes expression, and expression is where synthetic biology earns



its place. The strategy is to take the pathway out of its reluctant native context and rebuild it somewhere controllable. In practice that means one of two things: waking the cluster in its original host, or moving it into a heterologous chassis engineered to run it.

Activation in the native host uses the organism's own regulation as the lever. Overexpressing a pathway-specific activator, deleting a repressor, replacing a weak native promoter with a strong constitutive one, or perturbing global regulation through ribosome-engineering mutations can each shift a silent cluster into production ¹¹. CRISPR-based tools sharpened this considerably, allowing targeted promoter knock-ins and multiplexed edits in actinomycetes that were previously hard to manipulate ¹¹. The appeal is that the native host already possesses the precursors, cofactors and enzymatic context the pathway evolved with, so a product that folds and tailors correctly is more likely. The drawback is that many interesting producers remain genetically intractable, and native regulation can resist every lever at once.

Heterologous expression sidesteps the producer entirely. The cluster is captured, often by transformation-associated recombination or related direct-cloning methods that assemble large DNA fragments in yeast, then refactored, meaning its native promoters and regulatory elements are swapped for characterised parts, and introduced into a chassis such as a domesticated *Streptomyces* strain cleared of its own competing clusters, or into *Escherichia coli* or *Saccharomyces cerevisiae* ¹². Refactoring is

what converts a cluster from something a host merely carries into something it reliably transcribes. The genetic platforms for this, the capture vectors, the cleaned-up expression strains, and the libraries of insulated promoters, are now substantial ¹², and a recent synthesis frames synthetic biology explicitly as the route to revive microbial natural-product discovery rather than a peripheral convenience ¹³.

A parallel line of work attacks the uncultured problem from the other side, by making the unculturable grow. The isolation chip, a diffusion device that cultivates environmental bacteria in situ by letting native nutrients and signals pass through a membrane, allowed a previously uncultured soil organism to be grown and yielded teixobactin, an antibiotic that binds lipid-anchored cell-wall precursors and for which resistance was not readily selected in the laboratory ¹⁴. The result is a useful counterpoint to the genomic strategy. It shows that part of the hidden chemistry can be reached by better culturing rather than by moving DNA, and that the two approaches are complementary routes into the same dark biosynthetic space.

The proof that pathway engineering can reach the clinic predates the genomic framing. The semisynthetic artemisinin programme reconstructed a plant terpenoid pathway in yeast: an early strain produced the precursor artemisinic acid at titres far above any plant source ¹⁵, and a decade of further engineering raised production to the level where chemical conversion to artemisinin became an industrial reality ¹⁶. That work is the existence proof for the whole



enterprise. A complex, stereochemically demanding natural product, sourced for centuries from a crop, was rebuilt as a fermentation process through deliberate pathway engineering. If a plant sesquiterpenoid can be moved into yeast and optimised to commercial titre, a silent bacterial cluster can in principle be moved into *Streptomyces* and switched on.

In principle is the necessary qualifier. Expression remains the chronic bottleneck of the modern pipeline, and it is worth stating plainly rather than smoothing over. A refactored cluster may transcribe but fold or tailor its product incorrectly in a foreign host. Precursor and cofactor supply may be inadequate. The large nonribosomal peptide synthetase and polyketide synthase clusters, which encode some of the most valuable chemistry, are exactly the ones hardest to clone intact and express functionally. Where genome mining now generates candidate clusters faster than any laboratory can evaluate them, heterologous-expression throughput has not kept pace. This asymmetry, abundant predictions against scarce expression capacity, is the practical reason the field needs the third technology: if only a fraction of predicted clusters can ever be expressed, the choice of which to attempt has to be made well.

4. Artificial intelligence: prioritising clusters and molecules

Machine learning enters natural-product discovery on two distinct fronts, and conflating them has caused most of the overstatement in the area. The first front is genomic: predicting, from

sequence, which clusters exist and what they are likely to make or do. The second is chemical: predicting, from structure, which molecules are likely to be bioactive. They draw on different data and carry different risks, and they should be judged separately.

On the genomic front, the move is from fixed rules to learned patterns. DeepBGC treats a genome as a sequence of protein-family domains and uses a recurrent network to recognise the signature of a cluster, which lets it detect clusters whose architecture departs from the hand-written rules and reduces spurious calls¹⁰. Downstream classifiers then attempt the harder question of what a cluster does, assigning a predicted product class or a coarse activity label from cluster features, so that mining output can be ranked by predicted novelty and interest rather than read as an undifferentiated list¹⁰. This is prioritisation, and given the expression bottleneck described above, prioritisation is precisely the scarce commodity. A model that can tell a chemist which five of five hundred clusters to express first is doing the single most useful thing available.

The chemical front is where the headline results sit, and they are genuine. Stokes and colleagues trained a graph neural network on roughly 2,500 molecules labelled for growth inhibition of *E. coli*, then applied it to compound collections the model had never seen¹⁷. Screening the Drug Repurposing Hub of around 6,000 compounds, it flagged halicin, a molecule investigated years earlier for an unrelated indication, as a strong antibacterial with a structure unlike existing



antibiotics. Halicin proved active against *Clostridioides difficile*, *Acinetobacter baumannii* and *Mycobacterium tuberculosis*, and cleared a resistant *A. baumannii* infection in a mouse wound model¹⁷. The same model then scored more than 100 million molecules from a virtual library in a few days and nominated structurally novel candidates, a scale no physical screen approaches¹⁷. The lesson is not that a computer found an antibiotic. It is that a model generalised from a modest training set to chemistry it had never encountered.

The follow-up work answered the standing objection that a neural network is a black box offering no medicinal-chemistry rationale. Wong and colleagues built graph models that could be interrogated for the substructures driving their predictions, used that to extract a chemical rationale for antibacterial activity, and reported a new structural class active against methicillin-resistant *S. aureus*, with the model pointing to the specific substructures responsible¹⁸. Explainability is not a cosmetic feature here. A class defined by an understood substructure is one a chemist can optimise rationally and a regulator can reason about, which is a different and more durable proposition than a single lucky hit.

Machine learning also attacks the old dereplication problem from the metabolomics side, which closes the loop back to isolation. Global Natural Products Social Molecular Networking organises tandem mass spectra into networks in which structurally related molecules cluster together, so that known compounds

annotate by spectral match and the unmatched nodes, the candidate unknowns, stand out¹⁹. Layering learned spectral embeddings and substructure-prediction models on top of such networks pushes annotation further into uncharted chemistry¹⁹. The effect is to invert the economics that drove industry away in the first place: the molecules most likely to be already known are flagged early and set aside, concentrating scarce isolation effort on what is genuinely new.

Beyond classification, a third use is beginning to mature: generation and target reasoning. Generative models trained on known chemical space can propose scaffolds biased toward natural-product-like features, and models that predict a molecule's likely protein target offer a route from a mined structure to a mechanistic hypothesis before any assay is run. These are earlier in their evidence than the antibacterial-screening results and should be read as promising rather than proven, but they point at the same organising idea: at every stage where a human once chose by intuition, a model can now supply a ranked, testable shortlist. The value is not that the model is right, but that it is calibrated and fast enough to make triage affordable at the scale mining now produces.

The limits deserve the same prominence as the results, because they are routinely underplayed. Every supervised model here is bounded by its training data, and in natural-product space that data is scarce, imbalanced and biased toward well-studied chemistry and the easily cultured organisms that produced it. A model trained



mostly on *E. coli* growth inhibition generalises uncertainly to other pathogens and other mechanisms. Virtual screens of hundred-million-compound libraries return ranked lists, not drugs, and the attrition between a predicted hit and a validated lead, through synthesis, spectrum, toxicity, pharmacokinetics and resistance, remains severe and is rarely reported in the same breath as the hit rate. Benchmarking across tools is inconsistent, so published performance figures are hard to compare directly. None of this is a reason to discount the approach. It is a reason to read a model's output as a well-ranked hypothesis set rather than a result, which is how the strongest groups in fact use it ²⁰.

5. Convergence and critical appraisal

Treated separately, each technology has a characteristic failure. Genome mining predicts far more than can ever be expressed. Synthetic

biology can express a cluster but cannot, by itself, say which of thousands is worth the attempt. Machine learning ranks candidates but cannot make a molecule or prove its ranking without wet-laboratory confirmation. Set in sequence, each covers the others' weakness. Mining reads the biosynthetic potential of a genome or a metagenome; learning models rank the clusters by predicted novelty and the likely products by predicted activity; synthetic biology expresses the few that survive triage; and metabolomics with networked, learned annotation confirms the product and screens out rediscovery before isolation consumes any effort. This is the pipeline the strongest recent work is converging on (Figure 1), and the comprehensive survey of artificial intelligence for natural-product discovery describes essentially this integrated loop rather than a single decisive algorithm ²⁰.

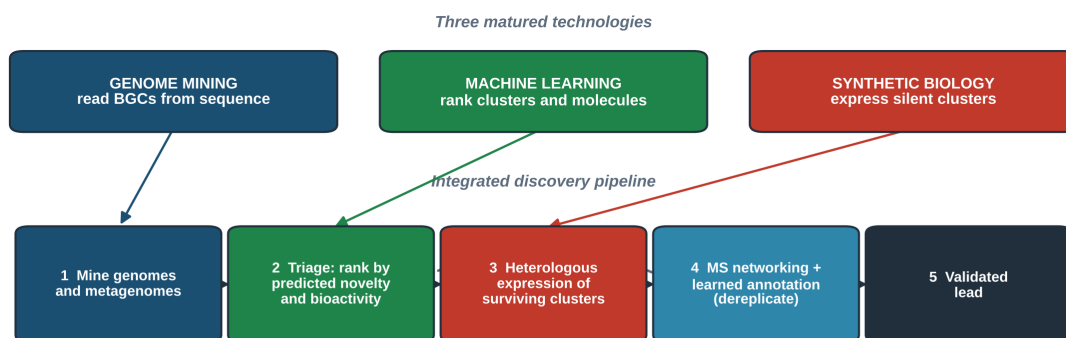


Figure 1. The three technologies and their convergence into one discovery pipeline. Genome mining, machine learning and synthetic biology matured largely independently (top) and now feed a single sequential workflow (centre): biosynthetic gene clusters (BGCs) are read from genomes and metagenomes (stage 1), ranked by predicted novelty and bioactivity (stage 2), expressed in tractable heterologous hosts (stage 3), and dereplicated by mass spectrometry (MS) networking with learned annotation before isolation (stage 4), yielding validated leads (stage 5); measured results can retrain the triage model (dashed return).

What the integration genuinely delivers is a change in where the funnel narrows. Classical discovery narrowed at the very start, bounded by

what would grow and what was abundant enough to isolate. The integrated pipeline narrows in the middle, at expression and computational triage,



both of which are improvable by engineering in a way that culturability was not. That is real progress, and it is why the underexploited chemical space of silent and uncultured biosynthesis is now addressable at all rather than merely lamented.

The appraisal has to stay honest about what is not yet solved, and the picture is uneven by design because the evidence is uneven. The expression bottleneck is the hard constraint. Until heterologous hosts become general rather than idiosyncratic, the pipeline will express only a minority of what it predicts, and the quality of computational triage will matter more than any further advance in detection sensitivity. The validation gap on the chemical side is the second constraint, and the literature's habit of foregrounding striking hits while leaving downstream attrition unquantified makes the whole approach look more finished than it is. Reproducibility and benchmarking are weak

across both the genomic and chemical tools, so claims of superiority between methods are often not comparable. And the training-data bias compounds itself in a way that is easy to miss: models learn from characterised chemistry, characterised chemistry comes from culturable organisms, and the uncultured majority that holds the real novelty is precisely where the models are least informed. A further tension runs through the whole enterprise and is rarely named. The economic logic that made natural products unattractive to industry, long timelines and uncertain returns, has not changed, and antibacterials in particular remain commercially punishing however they are discovered. A revived discovery engine produces candidates; it does not by itself fix the market that decides whether those candidates become drugs. None of this undercuts the central argument. It defines where the scientific work now lies, and marks the boundary beyond which the problem stops being technical.

Table 1. Representative technologies and milestones across the three pillars of revived natural-product discovery, with the contribution each makes to an integrated pipeline and its principal limitation. Abbreviations: BGC, biosynthetic gene cluster; MRSA, methicillin-resistant *Staphylococcus aureus*; MS, mass spectrometry.

Pillar	Representative tool or milestone	Contribution to the pipeline	Principal limitation
Genome mining	antiSMASH detection and classification ⁵	Reads BGCs and partial product structures from genome sequence	Detection does not establish expression; structure prediction reliable only for regular assembly lines
Genome mining	BiG-SCAPE cluster families and MIBiG reference set ^{7,8}	Ranks hits by novelty via similarity networks against characterised clusters	Depends on how completely known chemistry has been deposited
Genome mining	Evolution-guided discovery of corbomycin and complestatin ⁹	Targets novel mechanism, not merely novel structure	Requires a prior mechanistic hypothesis to guide the search
Synthetic	Heterologous expression	Expresses silent or	Chronic throughput



biology	and refactoring platforms ^{12,13}	uncultured clusters in tractable hosts	bottleneck; large clusters hard to clone and tailor correctly
Synthetic biology	In situ cultivation yielding teixobactin ¹⁴	Brings uncultured producers into culture without moving DNA	Applies only to organisms that can be coaxed to grow
Synthetic biology	Semisynthetic artemisinin in yeast ^{15,16}	Proves industrial-scale production of a complex natural product by pathway engineering	Each pathway still demands bespoke optimisation
Artificial intelligence	DeepBGC cluster prediction ¹⁰	Detects atypical clusters and ranks them by predicted novelty	Bounded by scarce, biased genomic training data
Artificial intelligence	Halicin and the explainable anti-MRSA class ^{17,18}	Predicts bioactivity and an actionable chemical rationale at scale	Hits are hypotheses; downstream attrition to a validated lead is large
Artificial intelligence	GNPS molecular networking with learned annotation ¹⁹	Dereplicates by MS so effort concentrates on genuine unknowns	Annotation confidence falls fastest exactly where chemistry is most novel

6. Future directions

The specific needs follow directly from those constraints rather than from a general call for more research. Expression must become general. The field needs a panel of well-characterised heterologous chassis with standardised, insulated genetic parts and reliable capture of large clusters, so that attempting a new cluster is a protocol rather than a research project, and so the design-build-test-learn cycle of strain engineering can be automated and fed by the computational triage upstream. Closing that loop, in which a model nominates clusters, an automated platform expresses them, and the measured results retrain the model, is the configuration most likely to produce compounding rather than one-off returns.

On the computational side, the priority is data and standards before architecture. Shared, honestly imbalanced benchmark sets for cluster detection, product-class prediction and bioactivity, with downstream attrition reported

alongside hit rates, would let the field compare methods and stop overclaiming. Models that integrate genomic, metabolomic and bioactivity data within one framework, rather than treating cluster prediction and molecule scoring as separate problems, are the natural next step, and the explainable approaches that yield a medicinal-chemistry rationale should be preferred over marginally more accurate black boxes, because an understood prediction is one a chemist can act on. Extending mining and expression reliably beyond actinomycetes into fungi, uncultured symbionts and complex environmental metagenomes is where the three-percent figure will actually be moved, since that is where the uncharacterised families are concentrated. Above all, the measure of success must shift from in silico hit counts and heterologous titres to prospectively validated leads carried into credible preclinical evaluation, because that endpoint is the only one that distinguishes a revived discovery engine from a productive source of hypotheses.



7. Conclusion

Natural products never stopped being the most productive source of validated drug chemistry; the machinery for reaching them had simply reached its limit. Genome mining has made the biosynthetic potential of a genome legible from sequence, synthetic biology has shown that a silent or foreign pathway can be rebuilt and switched on, and machine learning has shown that both clusters and molecules can be ranked well enough to matter, with halicin, a new anti-MRSA structural class and industrial artemisinin as concrete evidence rather than promise. The honest synthesis is that the bottleneck has moved rather than vanished: from finding novelty to expressing and prioritising it. The underexploited chemical space is now reachable in principle, and the task ahead is to make the reaching routine, which is an engineering and data problem the field is equipped to solve.

Declarations

Conflicts of interest: None declared.

Funding: This work received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions: All authors contributed to the conception, literature synthesis, drafting and critical revision of the manuscript, and approved the final version. Megha Verma is the corresponding author.

References

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod.* 2020;83(3):770-803. doi:10.1021/acs.jnatprod.9b01285.
2. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: advances and opportunities. *Nat Rev Drug Discov.* 2021;20(3):200-216. doi:10.1038/s41573-020-00114-z.
3. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell.* 2014;158(2):412-421. doi:10.1016/j.cell.2014.06.034.
4. Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat Microbiol.* 2022;7(5):726-735. doi:10.1038/s41564-022-01110-2.
5. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* 2023;51(W1):W46-W50. doi:10.1093/nar/gkad344.
6. Blin K, Kim HU, Medema MH, Weber T. Recent development of antiSMASH and other computational approaches to mine



- secondary metabolite biosynthetic gene clusters. *Brief Bioinform.* 2019;20(4):1103-1113. doi:10.1093/bib/bbx146.
7. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* 2020;48(D1):D454-D458. doi:10.1093/nar/gkz882.
8. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol.* 2020;16(1):60-68. doi:10.1038/s41589-019-0400-9.
9. Culp EJ, Waglechner N, Wang W, Fiebig-Comyn AA, Hsu YP, Koteva K, et al. Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature.* 2020;578(7796):582-587. doi:10.1038/s41586-020-1990-9.
10. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* 2019;47(18):e110. doi:10.1093/nar/gkz654.
11. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat Rev Microbiol.* 2015;13(8):509-523. doi:10.1038/nrmicro3496.
12. Zhang JJ, Tang X, Moore BS. Genetic platforms for heterologous expression of microbial natural products. *Nat Prod Rep.* 2019;36(9):1313-1332. doi:10.1039/c9np00025a.
13. Liu Z, Zhao Y, Huang C, Luo Y. Synthetic biology to revive microbial natural product discovery. *Nat Prod Rep.* 2024;41(5):803-822. doi:10.1039/d3np00057e.
14. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, et al. A new antibiotic kills pathogens without detectable resistance. *Nature.* 2015;517(7535):455-459. doi:10.1038/nature14098.
15. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature.* 2006;440(7086):940-943. doi:10.1038/nature04640.
16. Paddon CJ, Westfall PJ, Pitera DJ, Benjamin K, Fisher K, McPhee D, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature.* 2013;496(7446):528-532. doi:10.1038/nature12051.
17. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. *Cell.* 2020;181(2):475-483. doi:10.1016/j.cell.2020.01.021.
18. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with



explainable deep learning. Nature.

2024;626(7997):177-185.

doi:10.1038/s41586-023-06887-8.

19. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34(8):828-837. doi:10.1038/nbt.3597.
20. Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22(11):895-916. doi:10.1038/s41573-023-00774-7.