Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2025, Volume 5, Issue 2, Page No: 11-21

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Prioritizing Biosynthetic Gene Clusters When Genomes Overpredict Chemistry: Coupling Genomic Novelty, Metabolite Evidence, Ecological Context, and Chemical Tractability in Microbial Natural-Product Discovery

Ana Rodrigues[1]*, Tiago Martins[1], Bruno Lopes[2]

[1] Department of Biosynthetic Gene Cluster Prioritization, Faculty of Pharmacy, University of Lisbon, Lisbon, Portugal.
[2] Department of Genomic Novelty and Ecological Context, Faculty of Pharmacy, University of Porto, Porto, Portugal.

***E-mail** ✉ ana.rodrigues@ff.ulisboa.pt

Received: 21 April 2025; **Revised:** 12 August 2025; **Accepted:** 21 August 2025

ABSTRACT

Microbial genome mining routinely identifies far more biosynthetic gene clusters (BGCs) than can be experimentally characterized, creating a prioritization problem rather than a simple detection problem. Sequence novelty is frequently treated as the principal triage signal, yet sequence divergence does not guarantee production, structural novelty, ecological relevance, or practical chemical access. This article develops an original computational prioritization model that integrates four non-equivalent evidence domains: genomic novelty and biosynthetic interpretability, metabolite evidence, ecological context, and chemical tractability with rediscovery risk. The model is designed as evidence-aware decision support rather than a validated predictor of discovery success. It distinguishes encoded biosynthetic potential from observed chemistry, candidate BGC–metabolite associations from structural or causal assignments, and ecological plausibility from demonstrated function. It further treats tractability as an independent consideration because scientifically unusual clusters may remain silent, analytically masked, difficult to activate, or costly to isolate, whereas readily observable chemistry may be dominated by known products. The proposed architecture therefore favors convergent evidence, records provenance and uncertainty, and permits candidate priority to change as orthogonal evidence accumulates. Its intended use is to allocate experimental effort more transparently across microbial natural-product programs while preserving alternative explanations and failure states. The framework does not establish universal weights, calibrated probabilities of novel chemistry, or prospective superiority over existing workflows; those claims require class-diverse, taxonomically broad, prospective validation.

Keywords: Biosynthetic gene clusters, Genome mining, Metabolomics, Natural products, Ecological context, Dereplication

How to Cite This Article: Rodrigues A, Martins T, Lopes B. Prioritizing Biosynthetic Gene Clusters When Genomes Overpredict Chemistry: Coupling Genomic Novelty, Metabolite Evidence, Ecological Context, and Chemical Tractability in Microbial Natural-Product Discovery. Spec J Pharmacogn Phytochem Biotechnol. 2025;5(2):11-21. https://doi.org/10.51847/1wJbttBHL7

Introduction

Genome mining has transformed microbial natural-product discovery by revealing biosynthetic capacity that is not apparent from routine cultivation or metabolite profiling. Large-scale analyses of bacterial genomes and metagenome-assembled genomes show that encoded specialized-metabolite diversity substantially exceeds the fraction of chemistry that has been experimentally characterized [1]. This asymmetry changes the computational problem. Detecting another candidate locus is no longer sufficient justification for experimental pursuit when sequencing routinely returns many candidates per organism or community. The limiting resource becomes the capacity to decide which clusters merit cultivation changes, metabolomics, genetic intervention, heterologous expression, isolation, structure elucidation, or bioactivity testing. A useful prioritization strategy must therefore

separate evidence for existence of a biosynthetic locus from evidence that pursuing that locus is likely to produce informative chemistry.

Modern genome-mining platforms sharpen this distinction. antiSMASH 7.0 expands computational detection and annotation of BGCs, including predictions related to regulation and chemical features, but such output remains algorithmic evidence [2]. By contrast, MIBiG curates BGCs for which experimental characterization links genomic loci to known biosynthetic functions or products [3]. Prediction confidence and experimental validation are therefore not interchangeable. A close match to a characterized cluster can strengthen biosynthetic interpretation while simultaneously raising rediscovery concern; weak similarity to known clusters can increase exploration value while leaving product class, expression state, or recoverability uncertain. The same evidence can consequently increase one dimension of priority while decreasing another.

This article proposes that BGC selection should be framed as a multi-domain decision problem rather than a single novelty ranking. Four domains are kept analytically distinct: reference-relative genomic novelty and biosynthetic interpretability; evidence that corresponding chemistry is observed; ecological information that changes the plausibility or significance of expression; and chemical tractability, including dereplication burden and experimental accessibility. These domains are not assumed to contribute equally, nor are fixed universal weights proposed. Instead, each candidate is represented by an evidence profile whose components retain provenance, confidence, and boundary conditions.

The purpose is therefore not to replace existing genome-mining, metabolomics, or activation tools with another detector. It is to organize outputs from such tools into an explicit prioritization record that identifies why a BGC is being advanced, what remains unknown, and which next experiment would most reduce decision-relevant uncertainty. This structure also makes competing rationales visible: two clusters can receive similar overall attention for entirely different evidential reasons, and those reasons can imply different next experiments. The model is an original computational synthesis and remains a proposed decision architecture until evaluated prospectively across distinct taxa, BGC classes, analytical regimes, and discovery objectives.

The excess-prediction problem in genome mining

The scale of contemporary BGC resources illustrates why detection alone cannot function as a discovery endpoint. The antiSMASH database aggregates very large numbers of predicted BGC regions and provides sequence-based routes for comparing newly detected clusters with an expanding reference space [4]. Such growth is valuable for contextualization, but it should not be interpreted as proportional growth in experimentally observed metabolites. Database expansion reflects genome acquisition, assembly quality, annotation, and prediction behavior as well as genuine biosynthetic diversity. The relevant consequence for prioritization is a crowded candidate space in which raw presence carries progressively less discriminatory value.

Metagenomic resources further enlarge that space. BGC Atlas integrates biosynthetic predictions across extensive genomic and metagenomic collections with associated metadata, enabling global exploration of candidate diversity [5]. This scale makes novelty filters necessary but also exposes their limits. A cluster that is distant from familiar sequence space may be biologically interesting while remaining silent under tested conditions, fragmented in an assembly, experimentally inaccessible, or chemically convergent with known products. For this reason, the model developed here uses the term reference-relative genomic novelty rather than treating sequence distance as direct evidence of structural novelty.

Prediction architecture is another source of non-equivalence. Recent neural-network approaches can identify biosynthetic potential from learned sequence patterns [6], while DeepBGC established an earlier deep-learning strategy for BGC detection and product-class prediction [7]. Comparative frameworks such as BiG-SCAPE and CORASON instead organize relationships among clusters and gene-cluster families to explore biosynthetic diversity [8]. These approaches answer related but different questions, operate with different representations, and inherit different training, annotation, and boundary-definition limitations.

Accordingly, disagreement among prediction systems should not automatically be resolved by averaging scores, and agreement should not be treated as experimental confirmation. The proposed prioritization record instead preserves the origin of each genomic signal: detection method, family context, similarity basis, and reference coverage. This turns algorithmic multiplicity into provenance information and allows uncertainty to remain localized rather than hidden in a composite score. The aim is not to penalize computational prediction, but to

prevent a high-confidence sequence annotation from absorbing uncertainty that actually belongs to expression, chemistry, ecology, or experimental access.

What makes a biosynthetic gene cluster worth pursuing?

A BGC becomes worth pursuing when its evidence profile supports an informative experimental decision, not merely because it is unusual by one computational criterion. Targeted genome-mining approaches demonstrate that evolutionary or feature-specific context can reveal biosynthetic diversity that would be obscured by global similarity alone [9]. Such evidence can raise priority when it matches the discovery objective—for example, when a program seeks a particular biosynthetic capability or an underexplored evolutionary branch. It does not, however, establish that the cluster is expressed or that its product will be chemically novel.

Biosynthetic interpretability provides a second dimension. iPRESTO links recurrent biosynthetic subclusters to specific natural-product substructures, allowing local genomic organization to support chemically interpretable hypotheses [10]. This type of evidence can make a candidate more actionable because it suggests what structural element or biosynthetic transformation might be tested. Its resolution must remain explicit: a subcluster-to-substructure association does not identify the entire metabolite, and limited representation of characterized biosynthetic logic constrains inference for unusual pathways.

Validated reference knowledge is equally important for defining what "novel" means. MIBiG 4.0 expands curated experimentally validated BGC information and makes evidence provenance central to comparison [11]. Distance from characterized clusters can therefore contribute to novelty and rediscovery reasoning, but absence from a curated database cannot itself prove novelty. Reference resources are incomplete and inherit the taxonomic and chemical biases of what has already been studied.

The proposed criterion is consequently multidimensional. A candidate may be compelling because it occupies unusual genomic space, contains interpretable biosynthetic features, has orthogonal evidence of produced chemistry, arises in an informative ecological context, or presents an experimentally favorable route to characterization. These rationales should be judged against the discovery objective. A cluster sought for unprecedented scaffold exploration may tolerate weak initial observability, whereas a program constrained by rapid isolation may rationally favor existing metabolite evidence and lower experimental burden. Conversely, a strong signal in one domain can coexist with weakness in another. Priority should therefore reflect the value of resolving the candidate's remaining uncertainty, not an assumption that every desirable attribute must already be present. **Table 1** classifies the evidence states that should remain separate before later integration.

Table 1. Evidence classes that should remain distinct when judging BGC pursuit

Evidence class	Supports	Does not establish	Proposed role
Genomic novelty	Reference-relative sequence distinctiveness	Novel structure or production	Exploration value
Biosynthetic interpretability	Pathway or partial structural hypothesis	Complete product identity	Experimental actionability
Metabolite evidence	Detectable associated chemistry	Biosynthetic causality or exact structure	Production plausibility
Ecological context	Condition-, interaction-, or lineage-relevant plausibility	Demonstrated ecological function	Context-sensitive priority
Experimental validation	Stronger locus–product linkage	Universal transferability	Higher evidential confidence

Note: These non-equivalent classes are integrated without fixed universal weights or calibrated discovery probabilities.

Genomic, chemical, and ecological evidence

Genome sequence evidence becomes more informative when it can be connected to measured chemistry. Paired genomics–metabolomics resources, complementary scoring functions, and chemical-class matching can rank candidate associations between BGCs and metabolite features [12–14]. These approaches strengthen prioritization by asking whether genomic and chemical variation cohere across strains or samples, but the output remains an association until independently resolved. Co-occurrence, correlated abundance, or class compatibility can arise without direct biosynthetic causality; therefore, the proposed model records metabolite linkage as corroborating evidence rather than structural proof.

Feature-based molecular networking adds another analytical layer by organizing related MS/MS features into chemical neighborhoods [15]. Compared with genome prediction alone, an observed spectral feature provides stronger evidence that chemistry is produced under the measured condition. Yet spectral relatedness is not equivalent to exact molecular identity. Ionization behavior, acquisition settings, feature extraction, and library coverage can alter the observed network, so chemical observability and annotation confidence remain separate attributes.

The value of convergence is illustrated by integrated-omics discovery in which rule-based biosynthetic and chemical evidence supported recovery of antimicrobial macrocyclic peptides [16]. Such examples show that orthogonal evidence can convert a computational hypothesis into experimentally supported discovery evidence. They do not establish that the same combination of signals will perform equally across organisms, biosynthetic classes, or analytical workflows. The proposed model therefore rewards convergence conceptually without assigning an unvalidated universal multiplier.

Ecological evidence is similarly conditional. Environmental compartment, microbial interaction, and genomic background can influence specialized-metabolite expression or the interpretation of BGC evolution [17–19]. Ecological context should therefore modify priority only when the relevant condition or relationship is observed or otherwise justified; co-occurrence alone should not be translated into demonstrated ecological function.

Finally, underexplored host-associated lineages can possess distinctive natural-product repertoires, as shown for animal-associated marine Acidobacteria [20]. Such context can increase exploration value, especially when comparative evidence indicates biosynthetic distinctiveness. It does not guarantee laboratory expression, chemical novelty, or easy cultivation. The distinction matters because ecological rarity can justify exploratory effort while simultaneously increasing uncertainty about growth conditions, expression triggers, or analytical recovery. In the proposed architecture, ecological distinctiveness is therefore retained as an independent reason to investigate a candidate, not as a substitute for metabolite evidence or tractability.

Chemical tractability and rediscovery risk

A BGC can be scientifically compelling yet operationally unattractive if its chemistry is difficult to distinguish, observe, activate, isolate, or reproduce. Specialized metabolomics resources can reduce this burden by improving dereplication and focusing attention on features that are less readily explained by known chemistry. A myxobacterial natural-product database coupled to NMR-based metabolomics illustrates how taxon-focused chemical knowledge can support bioprospecting and reduce redundant characterization [21]. In the proposed model, however, failure to match a database entry is recorded as reduced rediscovery evidence, not proof of structural novelty. Coverage is uneven, and unmatched features can reflect missing reference chemistry, acquisition differences, or insufficient annotation.

Class-level annotation can add useful intermediate resolution. CANOPUS infers compound classes from high-resolution fragmentation spectra even when exact library matches are absent [22]. Such information can alter practical priority by indicating whether observed chemistry belongs to a familiar or underrepresented class, but class prediction does not establish exact molecular identity. Chemical evidence is therefore represented at the resolution actually achieved: observed feature, class-level assignment, candidate structure, or experimentally resolved product.

Tractability also includes the possibility that an attractive BGC is chemically invisible under routine conditions. Hidden biosynthetic potential may require altered cultivation, regulatory intervention, or other activation strategies before products become detectable [23]. Conversely, abundant common metabolites can mask less conspicuous chemistry; targeted inactivation of common-antibiotic clusters has exposed otherwise hidden antibiotics in actinomycetes [24]. These observations motivate a distinction between “not observed” and “not worth pursuing.” Absence of a metabolite signal may reflect silence, unsuitable conditions, analytical blind spots, pathway disruption, or a genuinely nonproductive prediction.

Unusual producers can further separate scientific value from practical feasibility. The Pendulisporaceae combine distinctive biology and specialized-metabolite potential with organism-specific experimental considerations [25]. Accordingly, tractability is modeled independently from novelty. A difficult producer need not be demoted if the scientific objective values unexplored chemistry sufficiently to justify intervention, whereas an easily cultivated producer may still warrant early deprioritization when strong dereplication evidence predicts little incremental

information. Tractability therefore represents the expected burden of obtaining discriminating evidence rather than a surrogate for intrinsic scientific worth.

Table 2 organizes the analytical and experimental constraints that can change whether a high-interest BGC is actionable.

Table 2. Analytical and experimental constraints that modify chemical tractability without defining scientific value

Constraint	Positive evidence contributes	Residual uncertainty	Proposed priority consequence
Dereplication context	Familiar versus weakly matched chemistry	Database coverage	Reduce redundant isolation effort
Class-level annotation	Chemical-family context	Exact structure unresolved	Refine follow-up strategy
Metabolite observability	Evidence of produced chemistry	Locus–product causality	Raise experimental actionability
Silent expression	Rationale for activation	Product may remain inaccessible	Add intervention cost
Masking by common products	Explanation for missed chemistry	Hidden product identity	Consider targeted unmasking
Producer tractability	Cultivation or manipulation feasibility	Transfer to scale uncertain	Separate feasibility from novelty

Note: Priority consequences are proposed decision interpretations, not validated quantitative penalties or bonuses.

The resulting logic is summarized in **Figure 1**, which preserves distinct evidence domains, decision gates, update paths, and the boundary between prioritization and validated discovery.

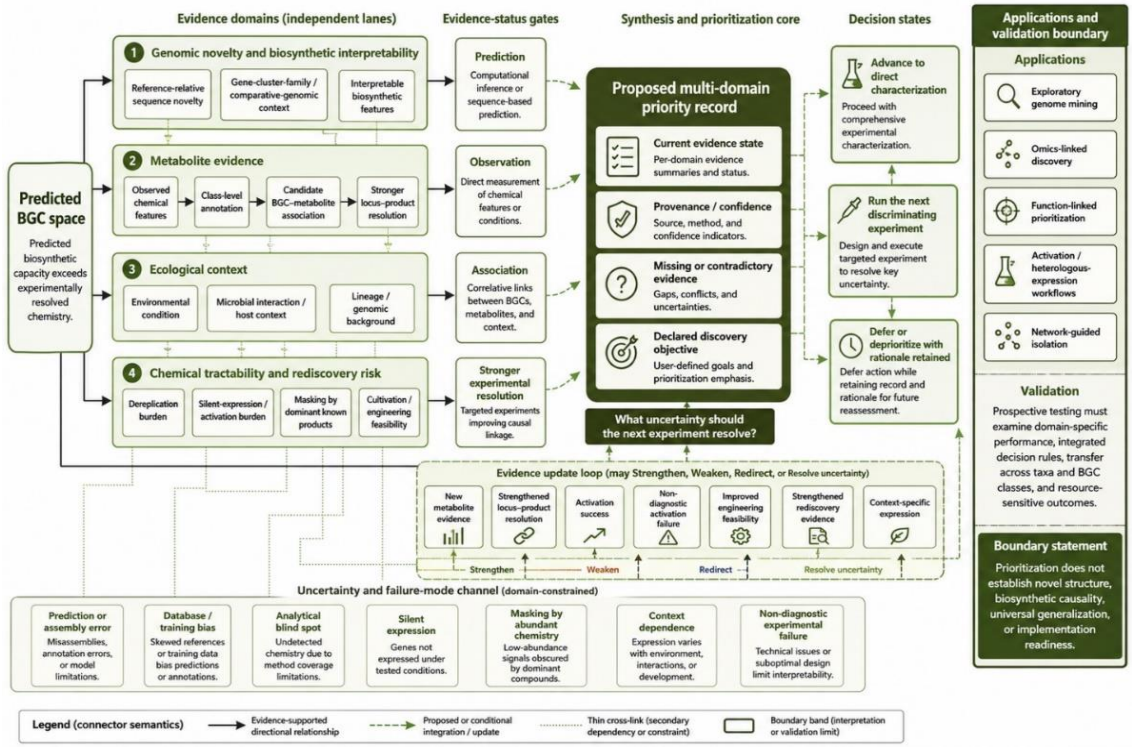

Figure 1. From predicted BGC space to revisable experimental priority: a proposed multi-domain decision architecture

A multi-domain prioritization model

The proposed model treats prioritization as structured evidence integration rather than as a universal scalar score. Each BGC is represented by four domain states: genomic evidence, metabolite evidence, ecological evidence, and

tractability/rediscovery evidence. Every state retains provenance, confidence, and missingness. The practical value of a domain depends on the discovery objective. A scaffold-diversity program may tolerate low initial observability for a genomically unusual candidate, whereas a rapid-isolation program may favor stronger metabolite evidence and lower activation burden. Consequently, missing information is not automatically scored as negative, and strength in one domain does not silently erase uncertainty in another.

Standardized comparative-genomic workflows provide one route for populating the genomic domain. BGCFlow supports systematic pangenome-scale analysis of BGCs across large genomic collections [26]. At broader scale, BiG-SLiCE and BiG-FAM organize gene-cluster-family relationships and searchable family space, enabling reference-relative placement of candidate clusters [27, 28]. These resources justify sequence-space features but not a calibrated probability that a chemically unprecedented molecule will be recovered. The model therefore stores genomic novelty as a bounded descriptor rather than translating distance directly into predicted chemical novelty.

Integration proceeds in two proposed stages. First, evidence-validity gates determine what each observation can legitimately support. For example, a spectral feature can establish chemical observability without establishing biosynthetic causality; an ecological association can change contextual plausibility without proving ecological function. Second, candidates are compared under a predeclared discovery objective using transparent decision rules. Suitable implementations could use ordinal domain states, Pareto-style comparison, or objective-specific lexicographic rules, but the present article does not claim that one mathematical implementation is universally optimal.

Missingness is represented explicitly rather than imputed as failure. “No metabolomics data,” “metabolomics collected but no corresponding feature detected,” and “a reproducible feature is present but remains unlinked” are computationally different states. The same principle applies to ecological metadata and tractability. Preserving these states prevents data-rich candidates from receiving an automatic advantage merely because more measurements have been performed and allows the model to distinguish an evidence deficit from evidence against a hypothesis.

Cross-modal evidence can materially improve actionability. Seq2PKS demonstrates that computational mass spectrometry and genome mining can narrow candidate relationships for type I cis-AT polyketides [29]. Likewise, self-resistance-guided mining can prioritize biologically relevant loci when that mechanism is informative for the product class [30]. These examples support modular, class-aware evidence features rather than mandatory criteria. A resistance signal may be highly useful for one discovery objective and irrelevant for another.

The core output is therefore a versioned priority record containing the candidate’s current rationale, contradictory evidence, missing domains, rediscovery concerns, and next discriminating experiment. The model is deliberately noncompensatory at critical boundaries: no amount of sequence novelty converts an unvalidated BGC–metabolite association into a causal assignment, and favorable tractability does not convert familiar chemistry into structural novelty. Its purpose is to make those boundaries computationally explicit while allowing objective-dependent comparisons among candidates that remain genuinely different.

Changing priority as new evidence appears

A useful prioritization model should change when the evidence changes. Static rankings are poorly matched to natural-product discovery because candidate status can shift after cultivation, metabolomics, genome editing, regulatory perturbation, or heterologous expression. Multi-pronged activation strategies in Actinomycetes demonstrate that experimental intervention can reveal chemistry that was not initially observable [31]. Within the proposed model, activation success therefore raises the evidence state for chemical observability and may justify further locus–product testing. It does not retrospectively prove that the initial high rank was correct, nor does it guarantee that the product is structurally novel or practically useful.

Evidence can also change tractability rather than scientific plausibility. Synthetic-biology approaches across the discover–design–build–test–learn cycle can expand options for pathway reconstruction, regulation, host choice, and production engineering [32]. Such information can move a candidate from experimentally inaccessible to actionable. The update should be attributed to improved intervention feasibility, not to increased genomic novelty.

Four proposed update directions are distinguished: strengthen, weaken, redirect, and resolve uncertainty. Orthogonal detection of a metabolite under a BGC-inducing condition may strengthen pursuit. Strong dereplication evidence for a common product may weaken a novelty-driven rationale. Failure of one activation

condition may redirect the next experiment without substantially lowering priority if that failure is non-diagnostic. Genetic or heterologous evidence that resolves a locus–product relationship can reduce uncertainty even if it does not improve tractability.

Each update should also be versioned. The computational record should retain the previous state, the new observation, the domain affected, the direction of change, and the reason the decision changed. This history prevents later successful experiments from obscuring earlier uncertainty and makes it possible to evaluate whether an initial prioritization rule was genuinely informative. It also distinguishes rank instability caused by accumulating biological evidence from instability caused by arbitrary changes in scoring conventions.

Priority updates need not be monotonic. Strong evidence of production can increase confidence while simultaneous dereplication reduces novelty-driven value. Improved heterologous-expression feasibility can raise tractability even while ecological relevance becomes less certain. The model therefore updates domains before updating the overall decision, preserving conflicting information instead of forcing every new datum into a positive or negative scalar increment.

Table 3 operationalizes these updates as reasoned state changes rather than numerical bonuses or penalties. This design also preserves negative results. A failed experiment is informative only to the extent that it tests a specified hypothesis under conditions capable of falsifying it; otherwise, the priority record should retain the failure without treating it as decisive evidence against the cluster.

Table 3. Proposed evidence-update rules for revising BGC priority during discovery

New evidence	Interpretation	Proposed priority update	Next discriminating action
Reproducible metabolite signal	Chemistry is observable	Strengthen actionability	Test locus–product linkage
Orthogonal locus–product evidence	Association uncertainty reduced	Strengthen confidence	Characterize structure/function
Activation succeeds	Prior silence was conditional	Raise tractability	Confirm product dependence
Activation fails once	Failure may be non-diagnostic	Retain or redirect	Change condition or strategy
Engineering route becomes feasible	Experimental access improves	Raise feasibility	Build/test pathway
Strong rediscovery match	Novelty rationale weakens	Deprioritize or redirect	Seek differentiating chemistry
Context-specific expression	Priority becomes condition-dependent	Refine context	Replicate relevant condition

Note: Updates are proposed qualitative state changes. No universal numerical increment, penalty, or decision threshold is implied.

Applications in microbial natural-product discovery

The proposed model is most useful when the discovery objective is explicit before ranking begins. In untargeted exploration, reference-relative genomic novelty and underexplored ecological context may dominate early triage, while the next experiment is chosen to establish observability. In metabolomics-rich collections, chemical evidence can instead identify candidates whose genomic and spectral patterns converge, shifting resources toward locus–product resolution. These are different decision problems and should not be forced into one fixed weighting scheme.

Integrated discovery studies illustrate several application routes. Rule-based omics mining has shown how biosynthetic and chemical evidence can converge on experimentally recovered bioactive macrocyclic peptides [16]. For a structurally constrained biosynthetic class, Seq2PKS shows how computational mass spectrometry can narrow polyketide candidates before costly characterization [29]. Self-resistance-guided mining provides a function-linked route for prioritizing relevant *Streptomyces* loci when resistance biology is informative [30]. Network-guided isolation coupled to heterologous production has likewise connected genome mining with recoverable macrolactam chemistry [33]. Each application demonstrates a different evidential path; none establishes a universal ranking formula.

The model can therefore serve as a project-level decision layer above existing tools. A microbial collection could first annotate BGCs, locate them in reference space, attach available metabolomics and ecological metadata, and

record practical constraints. Candidates would then be sorted by the declared objective and by the value of the next discriminating experiment. For a silent but highly distinctive cluster, the next action may be activation. For an observed but poorly assigned metabolite family, it may be genetic linkage. For a chemically familiar profile, it may be early dereplication rather than isolation.

Resource constraints can also become explicit without being confused with biological evidence. When two candidates have comparable scientific rationale, the model can favor the experiment expected to resolve the more consequential uncertainty with the available cultivation, analytical, or engineering capacity. Conversely, a resource-intensive candidate can remain strategically important when its evidence profile addresses a discovery objective that easier candidates cannot satisfy.

This structure is also suited to portfolio management. Rather than advancing only the highest nominal rank, a program can maintain candidates representing different evidence profiles and risk levels. Such diversification protects against systematic failure of a single evidence domain and makes resource allocation auditable. Importantly, "application" here means decision support within discovery workflows. It does not imply that the proposed architecture has been prospectively shown to increase hit rate, chemical novelty, productivity, or translational success.

Limitations and validation

The proposed model inherits uncertainty from every component it integrates. BGC prediction depends on model architecture, training representation, sequence context, and cluster boundaries [6]. Deep-learning approaches similarly remain constrained by the data and annotations from which they learn [7]. Correlation- and class-based genome–metabolome linking can rank plausible associations without establishing causality [14], while class-level spectral inference does not resolve exact molecular structures [22]. Integration cannot remove these limitations; it can only prevent them from being hidden.

Context introduces additional instability. Environmental expression patterns can differ across microbial compartments [17]. Interspecies interactions can reshape specialized metabolism [18]. Genomic or ecological background can change how BGC evolution is interpreted [19]. Laboratory activation further creates intervention-specific states that may not reproduce native ecological regulation [31]. Context must therefore be recorded as provenance, not converted into a universal bonus.

Prospective validation should test both the individual domains and the integrated decision rules. Curated BGC repositories provide essential references but remain incomplete [11]. Successful integrated-omics discovery demonstrates feasibility of particular evidence combinations rather than validation of the proposed model [16]. Class-specific workflows such as Seq2PKS likewise should not be generalized beyond their validated scope [29]. Successful heterologous production does not establish universal implementation readiness [33].

A rigorous evaluation would freeze the ranking logic before candidate selection, define success and failure outcomes in advance, compare domain ablations, and test transfer across taxa, BGC classes, data-availability regimes, and discovery objectives. Validation should distinguish rank discrimination from calibration, decision utility, stability under missing data, and experimental resource use rather than reporting only whether a selected candidate eventually produced an interesting metabolite. Held-out taxonomic or biosynthetic classes would be particularly important for detecting dependence on familiar reference space. Failure analysis should separately record inaccurate prediction, nonproduction, analytical nondetection, rediscovery, unsuccessful intervention, and unresolved attribution. Until such studies are performed, the framework should be interpreted as a falsifiable computational prioritization proposal rather than a validated predictor.

Conclusion

Microbial genome mining has made biosynthetic prediction abundant, but experimental attention remains scarce. The resulting challenge is not simply to identify more BGCs; it is to decide which candidates justify the next unit of experimental effort and why. The model proposed here addresses that problem by keeping genomic novelty, metabolite evidence, ecological context, and chemical tractability analytically distinct before integrating them into a transparent, revisable priority record.

This separation is essential because the evidence domains answer different questions. Sequence divergence can motivate exploration without proving novel chemistry. An observed spectral feature can establish chemical

observability without establishing biosynthetic causality or complete molecular identity. Ecological context can change plausibility without demonstrating function. Tractability can determine whether a valuable hypothesis is experimentally actionable without defining its scientific importance. A prioritization system that collapses these distinctions into an opaque score risks converting uncertainty into false precision.

The proposed architecture instead emphasizes evidence provenance, explicit discovery objectives, conditional decision gates, missingness, and updates when new information appears. It also preserves negative and contradictory evidence so that deprioritization remains explainable rather than arbitrary. Its principal practical contribution is therefore a disciplined way to connect genome-mining abundance to experimentally discriminating choices while retaining the reason each candidate was advanced, deferred, or redirected.

The framework remains a proposal. It does not establish universal weights, prospective predictive performance, or superiority over existing discovery workflows. Those questions require independent, prospective evaluation across diverse organisms, biosynthetic classes, analytical platforms, ecological settings, and experimental constraints. Such validation would determine whether evidence-aware prioritization can reliably convert genomic abundance into more efficient and informative natural-product discovery.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. 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):726-35. doi:10.1038/s41564-022-01110-2
2. Blin K, Shaw S, Augustijn HE, Medema MH, Weber T. antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023;51(W1):W46-50. doi:10.1093/nar/gkad344
3. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, et al. MIBiG 3.0: A community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 2023;51(D1):D603-10. doi:10.1093/nar/gkac1049
4. Blin K, Shaw S, Medema MH, Weber T. The antiSMASH database version 4: Additional genomes and BGCs, new sequence-based searches and more. Nucleic Acids Res. 2024;52(D1):D586-9. doi:10.1093/nar/gkad984
5. Bağcı C, Nuhamunada M, Goyat H, Ladanyi C, Sehnal L, Blin K, et al. BGC Atlas: A web resource for exploring the global chemical diversity encoded in bacterial genomes. Nucleic Acids Res. 2025;53(D1):D618-24. doi:10.1093/nar/gkae953
6. Lai Q, Yao S, Zha Y, Zhang HH, Zhang HB, Ye Y, et al. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model. Nucleic Acids Res. 2025;53(7):gkaf305. doi:10.1093/nar/gkaf305
7. 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
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):60-8. doi:10.1038/s41589-019-0400-9
9. Cediel-Becerra JDD, Cumsille A, Guerra S, Ding Y, de Crécy-Lagard V, Chevrette MG. Targeted genome mining with GATOR-GC maps the evolutionary landscape of biosynthetic diversity. Nucleic Acids Res. 2025;53(13):gkaf606. doi:10.1093/nar/gkaf606

10. Louwen JJR, Kautsar SA, van der Burg S, Medema MH, van der Hooft JJJ. iPRESTO: Automated discovery of biosynthetic sub-clusters linked to specific natural product substructures. PLoS Comput Biol. 2023;19(2):e1010462. doi:10.1371/journal.pcbi.1010462
11. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, et al. MIBiG 4.0: Advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res. 2025;53(D1):D678-90. doi:10.1093/nar/gkae1115
12. Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, et al. A community resource for paired genomic and metabolomic data mining. Nat Chem Biol. 2021;17(4):363-8. doi:10.1038/s41589-020-00724-z
13. Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLoS Comput Biol. 2021;17(5):e1008920. doi:10.1371/journal.pcbi.1008920
14. Louwen JJR, Medema MH, van der Hooft JJJ. Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. Microbiome. 2023;11(1):13. doi:10.1186/s40168-022-01444-3
15. Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. Nat Methods. 2020;17(9):905-8. doi:10.1038/s41592-020-0933-6
16. Cheng Z, He BB, Lei K, Gao Y, Shi Y, Zhong Z, et al. Rule-based omics mining reveals antimicrobial macrocyclic peptides against drug-resistant clinical isolates. Nat Commun. 2024;15(1):4901. doi:10.1038/s41467-024-49215-y
17. Geller-McGrath D, Mara P, Taylor GT, Suter EA, Edgcomb V, Pachiadaki M. Diverse secondary metabolites are expressed in particle-associated and free-living microorganisms of the permanently anoxic Cariaco Basin. Nat Commun. 2023;14(1):656. doi:10.1038/s41467-023-36026-w
18. Krespach MKC, Stroe MC, Netzker T, Rosin M, Zehner LM, Komor AJ, et al. Streptomyces polyketides mediate bacteria–fungi interactions across soil environments. Nat Microbiol. 2023;8(7):1348-61. doi:10.1038/s41564-023-01382-2
19. Salamzade R, Kalan LR. Context matters: Assessing the impacts of genomic background and ecology on microbial biosynthetic gene cluster evolution. mSystems. 2025;10(3):e01538-24. doi:10.1128/msystems.01538-24
20. Leopold-Messer S, Chepkirui C, Mabesoone MFJ, Meyer J, Paoli L, Sunagawa S, et al. Animal-associated marine Acidobacteria with a rich natural-product repertoire. Chem. 2023;9(12):3696-713. doi:10.1016/j.chempr.2023.11.003
21. Wang DG, Wang CY, Hu JQ, Wang JJ, Liu WC, Zhang WJ, et al. Constructing a myxobacterial natural product database to facilitate NMR-based metabolomics bioprospecting of myxobacteria. Anal Chem. 2023;95(12):5256-66. doi:10.1021/acs.analchem.2c05145
22. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39(4):462-71. doi:10.1038/s41587-020-0740-8
23. Scherlach K, Hertweck C. Mining and unearthing hidden biosynthetic potential. Nat Commun. 2021;12(1):3864. doi:10.1038/s41467-021-24133-5
24. Culp EJ, Yim G, Waglechner N, Wang W, Pawlowski AC, Wright GD. Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. Nat Biotechnol. 2019;37(10):1149-54. doi:10.1038/s41587-019-0241-9
25. Garcia R, Popoff A, Bader CD, Löhr J, Walesch S, Walt C, et al. Discovery of the Pendulisporaceae: An extremotolerant myxobacterial family with distinct sporulation behavior and prolific specialized metabolism. Chem. 2024;10(8):2518-37. doi:10.1016/j.chempr.2024.04.019
26. Nuhamunada M, Mohite OS, Phaneuf PV, Palsson BO, Weber T. BGCFlow: Systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. Nucleic Acids Res. 2024;52(10):5478-95. doi:10.1093/nar/gkae314

27. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. Gigascience. 2021;10(1):giaa154. doi:10.1093/gigascience/giaa154
28. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: The biosynthetic gene cluster families database. Nucleic Acids Res. 2021;49(D1):D490-7. doi:10.1093/nar/gkaa812
29. Yan D, Zhou M, Adduri A, Zhuang Y, Guler M, Liu S, et al. Discovering type I cis-AT polyketides through computational mass spectrometry and genome mining with Seq2PKS. Nat Commun. 2024;15(1):5356. doi:10.1038/s41467-024-49587-1
30. Yuan Y, Huang C, Singh N, Xun G, Zhao H. Self-resistance-gene-guided, high-throughput automated genome mining of bioactive natural products from Streptomyces. Cell Syst. 2025;16(3):101237. doi:10.1016/j.cels.2025.101237
31. Tay DWP, Tan LL, Heng E, Zulkarnain N, Ching KC, Wibowo M, et al. Exploring a general multi-pronged activation strategy for natural product discovery in actinomycetes. Commun Biol. 2024;7(1):50. doi:10.1038/s42003-023-05648-7
32. Foldi J, Connolly J, Takano E, Breitling R. Synthetic biology of natural products engineering: Recent advances across the Discover–Design–Build–Test–Learn cycle. ACS Synth Biol. 2024;13(9):2684-92. doi:10.1021/acssynbio.4c00391
33. Seibel E, Um S, Dayras M, Bodawatta KH, de Kruijff M, Jønsson KA, et al. Genome mining for macrolactam-encoding gene clusters allowed for the network-guided isolation of β-amino acid-containing cyclic derivatives and heterologous production of ciromicin A. Commun Chem. 2023;6(1):257. doi:10.1038/s42004-023-01034-w