

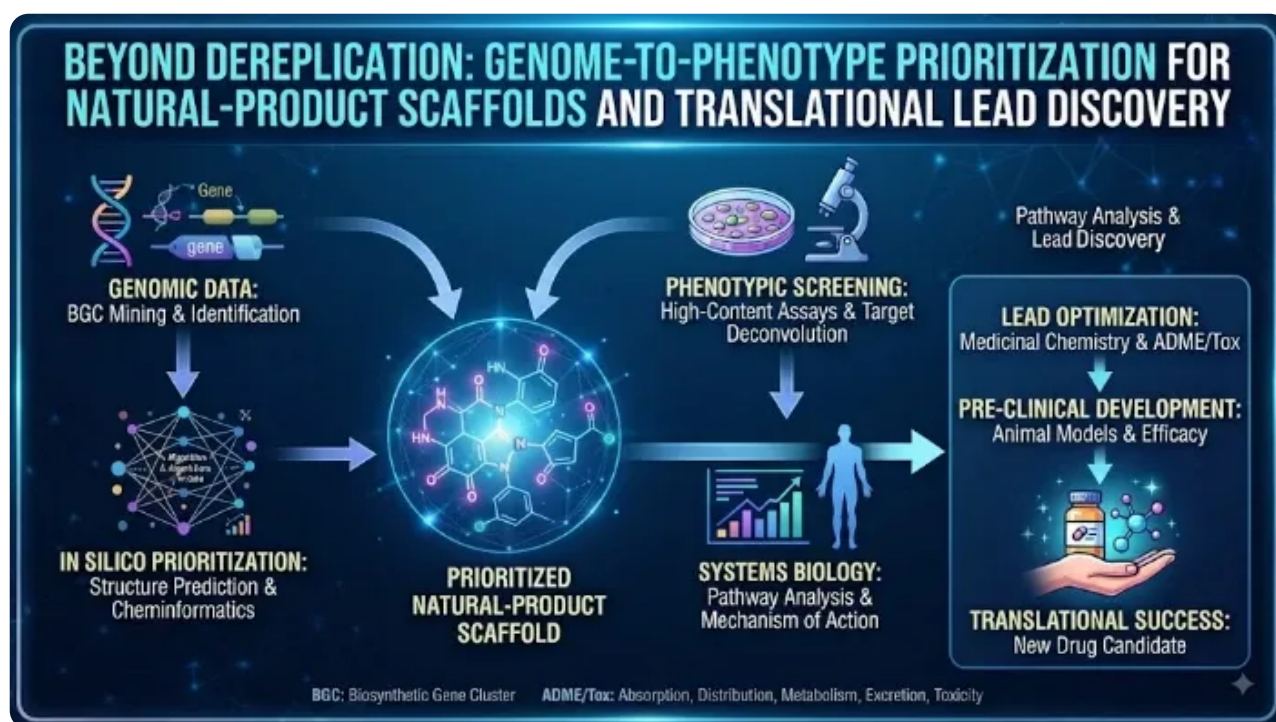
Beyond Dereplication: Genome-to-Phenotype Prioritization for Natural-Product Scaffolds and Translational Lead Discovery

Md. Humayan Kabir

May 31, 2026

Abstract

This article moves natural-product discovery beyond simple dereplication by proposing genome-to-phenotype prioritization. It explains how genome sequencing can reveal biosynthetic gene clusters that represent far more chemical potential than is observed under standard culture conditions, but also stresses that a predicted cluster is not automatically a useful drug lead. Metabolomics, gene deletion, isotopic labeling, heterologous expression, pathway activation, and phenotype screening are presented as experimental bridges linking biosynthetic genes to actual chemical scaffolds and biological activity. The article emphasizes prioritizing strains and clusters using multiple dimensions: phylogenetic distance, biosynthetic diversity, metabolomic profiles, phenotype strength, mechanistic novelty, and translational tractability. Culture conditions, co-culture, environmental stress, and engineered hosts are discussed as ways to activate silent pathways. Machine learning can integrate sequence novelty, predicted chemistry, metabolomic evidence, and biological scores, but uncertainty must remain visible. Ultimately, the workflow ranks natural-product opportunities by genetic novelty, expression evidence, phenotype, mechanism, supply feasibility, developability, ecological context, and scalable production potential.



A paradox has long been the problem with natural product discovery. The diversity of chemicals produced by microbes, plants, fungi, and marine organisms is extraordinary, and screening programs routinely overlook welcome re-discoveries of old molecules, and huge genomic capacity is unexpressed and uncharacterized. Dereplication is useful to eliminate known compounds, but it is by no means sufficient. A better strategy would involve associating biosynthetic gene(s) with the observed phenotypes in such a way that natural-product scaffolds would be prioritized according to their biological relevance, novelty, and their translational potential, but not only on the basis of novelty.

One of the greatest advances in natural product discovery was the application of genome sequencing to the identification of biosynthetic genes clusters, or BGCs, that code for secondary metabolite pathways, on the front end of the discovery process. These clusters can be identified and compared to other clusters in large genome collections with the help of tools like antiSMASH, BiG-SCAPE and curated resources like MIBiG. The resulting maps reveal that many organisms have a much greater ability for biosynthesis than are observed in standard culture conditions. The information in the sequence thus reveals a secret chemical stockpile.

The issue is, a projected BGC is not a drug lead. Some clusters are non-silent, weak expression, condition-dependent, fragment or misassigned. Some encode close analogues of known chemistry, but are seemingly different at the sequence level. In order to discover the connections between a cluster and an actual molecule, and then between the molecule

and its phenotype, experimental bridges would need to be built.

One such bridge is provided by metabolomics. Metabolomics can be used to compare the features of metabolites between strains, culture conditions, gene perturbations, engineered hosts, etc. Once a candidate relationship has been correlated with a BGC, it is correlated with the presence of a chemical feature. Strong causal evidence that a specific gene cluster is responsible for a specific scaffold is then provided by isotopic labeling, targeted gene deletion, and heterologous expression, and activation of pathways.

Another layer of prioritization is added by the Phenotype. The chemistry is given biological meaning by its antimicrobial activity, by its ability to selectively inhibit certain cancer cells, by its effect on the immune system, by its enzyme inhibitory activity, or by some other functional association. A genome to-phenotype workflow is a series of steps in which the BGCs that yield reproducible activity are selected, and then the metabolite responsible is identified. This prevents the same amount of effort being put into each genetically novel cluster and focuses resources by concentrating on biological function as it is seen.

Under this system, the selection of strains is more complex. Redundant chemistry is not reliably predicted by phylogeny, as clusters that are present in both strains might be expressed differently. A biosynthetic diversity has been shown to exist within *Streptomyces* collections to a high degree, depending upon the strain. Prioritization is therefore more useful if done based on phylogenetic distance, BGC content, metabolomic profiles and phenotype, instead of relying on one of these as a short-cut.

The culture conditions can also serve as experimental variables. Pathways may be activated under normal lab growth conditions, but when the culture conditions are changed, including nutrient composition, salinity, pH, temperature, co-culture, stress, and signaling molecules, they may be activated. This is the “one strain, many compounds” concept of plasticity. Systematic condition matrices associated to transcriptomics and metabolomics assist decide if a potential BGC is obtainable for fermentation or should be immediately genetically reconstructed.

When the amount of native production is limited or the species cannot be genetically manipulated, heterologous expression can be useful. Whole clusters of genes can be isolated and expressed in more well-characterized hosts with the promoters, precursor supply and regulatory elements being optimized. This approach will address the pathway potential while eliminating some of the constraints of the original organism. It also provides a working path towards scalable production if the scaffold is a significant lead.

Machine learning can help with the integration of sequence novelty, domain architecture, predicted chemistry, metabolomic correlations, and phenotypic scores to aid in prioritization. The challenge is: relatively low number of BGCs that have tested their products and mechanisms. Uncertainty is thus required to be transparent and validation is required in models. The high novel score is not a justification for experiments, but rather a guideline for them.

What is new about the chemistry? A novel scaffold is exciting, but is also important that it has translational value, be potent, selective, have appropriate physicochemical properties, can be synthesized or fermented, and has a mechanism. In some cases the complexity of the structure leads to the identification of new biological targets even if direct development is not possible. Others offer starting points for semisynthetic simplification or for medicinal-chemistry optimization.

Mechanism should be introduced later after the phenotypic confirmation. The identification of targets or pathways is achieved by the help of resistant mutants, chemoproteomics, genetic screens, biochemical assays and omics responses. Linking mechanism back to the BGC allows for better prioritization since a cluster encoding chemistry with a different mechanism should be prioritized over another analogue with a similar mechanism, but a target class that is already crowded. Discovery of natural products is not just for novelty but mechanism-aware.

The dereplication step is still very much needed in the entire process. Known metabolites can be identified prior to extensive purification with the aid of spectral libraries, molecular networking, accurate mass, isotope patterns, NMR databases and literature comparison. Dereplication is best applied as an ongoing rather than a 'first cut' filtering operation. A new activity may result from a familiar compound synthesized under an unusual set of circumstances, and the quick identification of the new compound saves resources.

In addition to supply, translational prioritization takes into account other factors. The different challenge for development is the case of a microgram producer as opposed to a strong microbial producer. Pathway genes can be transferred, optimized or engineered with the help of genome information. Knowledge of the above precursor requirements, enzyme complexity and the regulation of the pathways helps to make realistic choices of fermentation, total synthesis or semisynthesis.

Data standards bolster the entire data pipeline. The more information that is reported consistently and experimentally validated BGC databases gain in value, especially when it

comes to the structures, genes, bioactivities, production conditions, and taxonomic information. Standardization facilitates benchmarking of the prediction tools and eliminates the re-discovery of them. It also generates training data for models which try to relate sequence features to chemical classes or biological functions.

Thus, a practical prioritization score must have multiple dimensions: novel genetics, BGC confidence, evidence of expression, metabolite uniqueness, strength of phenotype, mechanistic differentiation, tractable supply and early developability. Success isn't measured by just one criterion. Is to make trade offs clear, so that the eye-catching genome mining result does not outrank the slightly less novel cluster with much greater translatable potential.

The ecological context can be used to further enhance prioritisation. This metabolite may represent a biological function specific to microbial interrelationships, symbiosis, limiting nutrients or host association that is not detected by standard fermentation. When genome and metabolite information is paired with environmental data, clues are provided as to the natural activation of a pathway, which can be complementary. This information can inform the design of culture and serve to reinforce the hypotheses of antimicrobial, signaling, or defensive roles.

Other emerging layers are single-cell and spatial approaches. Mixed microbial communities can also harbor rare producers whose chemistry is lost in bulk isolation and spatial metabolomics can be used to link a molecular feature to a given region of colony or interaction. These patterns can be correlated to metagenomic BGCs, which have contributed to the genome-to-phenotype discovery that goes beyond easily cultured strains. This extends the range of accessible chemical diversity, while maintaining the mechanistic attribution.

Intellectual-property strategy is also included in translational prioritization. While novelty is apparent with a completely new scaffold, a scaffold known to exist with a new mechanism, production route or optimized derivative can also be valuable. With advance landscape analysis, scientific teams are not given free rein to develop simply for the novelty of the detection. Along with the biological and manufacturing evidence, there is a new filter that is considered – commercial.

The most experimental humble prioritization systems are the best. For complex biology, sequence similarity, spectral networking and computational scoring all pack the problem into a manageable list of ranked hits, but none of them is a solution in isolation to the problem of having to isolate, structure and test the function of each. The transparent

ranking should include the explanation of how a candidate got to the top and what uncertainty exists. That way the abundance, measured by the tools of bioinformatics, is transformed into a series of smart and testable experiments.

Natural-product discovery is not just dereplication, it's a string of information layers. Biosynthetic potential is inspected through the genome; expressed chemistry is identified by metabolomics; cluster-product relationships are established by genetics; biological meaning is determined by function, which is revealed by phenotype; and mechanism is determined by mechanistic assays. Then, with a focus on supply and developability, the filters for translation ask whether more investment is warranted. By using the genome-to-phenotype prioritization, an immense number of predicted natural products is reduced to a focus on scaffolds that have a potential path towards useful medicines.