Nerpa 2: linking biosynthetic gene clusters to nonribosomal peptide structures

Ilia Olkhovskii,[1,2,3] Aleksandra Kushnareva,[1,2] Petr Popov,[4] Azat Tagirdzhanov[1,2] and Alexey Gurevich[1,2,*]

[1] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbrücken 66123, Germany, [2]Center for Bioinformatics Saar and Saarland University, Saarland Informatics Campus, Saarbrücken, 66123, Germany, [3]Saarbrücken Graduate School of Computer Science, Saarland University, Saarbrücken 66123, Germany and [4]Neapolis University Pafos, Pafos 8042, Cyprus

*Corresponding author. Email: alexey.gurevich@helmholtz-hips.de

Abstract

Summary: Nonribosomal peptides (NRPs) are medically important natural compounds. Linking biosynthetic gene clusters (BGCs) encoding NRP synthesis in microbial genomes to their final products is challenging due to the modularity, promiscuity, and post-synthesis modifications that contribute to NRP diversity. Here, we present Nerpa 2, a major update to our previous tool for linking computationally predicted BGCs to a database of known NRP chemical structures. Nerpa 2 incorporates substantial software and algorithmic improvements, including a new probabilistic alignment and scoring model. These advancements, combined with an extended training dataset, significantly improve the tool's accuracy and sensitivity. With its comprehensive and user-friendly HTML reports, Nerpa 2 is a powerful addition to the natural product research toolkit, accelerating NRP discovery.

Availability and implementation: Nerpa 2 is freely available at https://github.com/gurevichlab/nerpa.

1. Introduction

Nonribosomal peptides (NRPs) are a promising source of bioactive compounds with high pharmaceutical potential (Agrawal *et al.*, 2017; Süssmuth and Mainz, 2017). Their biosynthesis is governed by specialized enzymes encoded in biosynthetic gene clusters (BGCs). Linking BGCs to their products helps researchers predict and manipulate the production of these compounds, potentially leading to the discovery of new therapeutic agents (Medema *et al.*, 2021). The modern genome mining methods enabled the collection of hundreds of thousands of computationally predicted BGCs encoding unknown NRPs (Bağcı *et al.*, 2024). At the same time, specialized databases like Norine (Flissi *et al.*, 2020) contain thousands of NRP chemical structures. However, peculiarities in the NRP biosynthesis make the linking between these two types of data extremely difficult. As a result, only a small fraction of currently elucidated NRPs are connected to their respective BGCs (Dejong *et al.*, 2016; Flissi *et al.*, 2020; Zdouc *et al.*, 2025).

BGCs encoding NRPs have a modular structure that acts as an assembly line, where each module incorporates a single amino acid to the growing peptide chain (Süssmuth and Mainz, 2017). Modules, in turn, consist of domains, each having a specific function. The key domain defining the module specificity is the Adenylation (A) domain, while some others can introduce downstream modifications to the selected amino acid, such as methylation or epimerization. Some A domains are promiscuous and can attract various building blocks, enabling a single BGC to encode several closely related NRPs. Further, the gene activation order during the NRP biosynthesis is not always in line with the gene order in the corresponding BGC, leading to so-called *non-collinear* assembly lines. They might involve gene rearrangement, skipping, stuttering, and iterative reuse (Süssmuth and Mainz, 2017; Juguet *et al.*, 2009). Finally, a linear peptide sequence produced by an assembly line often undergoes macrocyclization or other tailoring reactions, resulting in a complex structure of the mature NRP product (Sieber and Marahiel, 2003).

Earlier, we developed Nerpa, a tool for high-throughput linking of BGCs to their tentative products from a database of NRP chemical structures (Kunyavskaya *et al.*, 2021). The method relied on the leading genome mining and NRP retrobiosynthesis software (Blin *et al.*, 2019; Ricart *et al.*, 2019) and the knowledge-based alignment procedure utilizing the classical Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). We trained the tool parameters using known BGC-NRP pairs from MIBiG, the largest database of experimentally verified BGCs (Kautsar *et al.*, 2020). Altogether, this allowed Nerpa to outperform the only competing method, GARLIC (Dejong *et al.*, 2016), in computational efficiency and accuracy. However, Nerpa still struggled to handle many

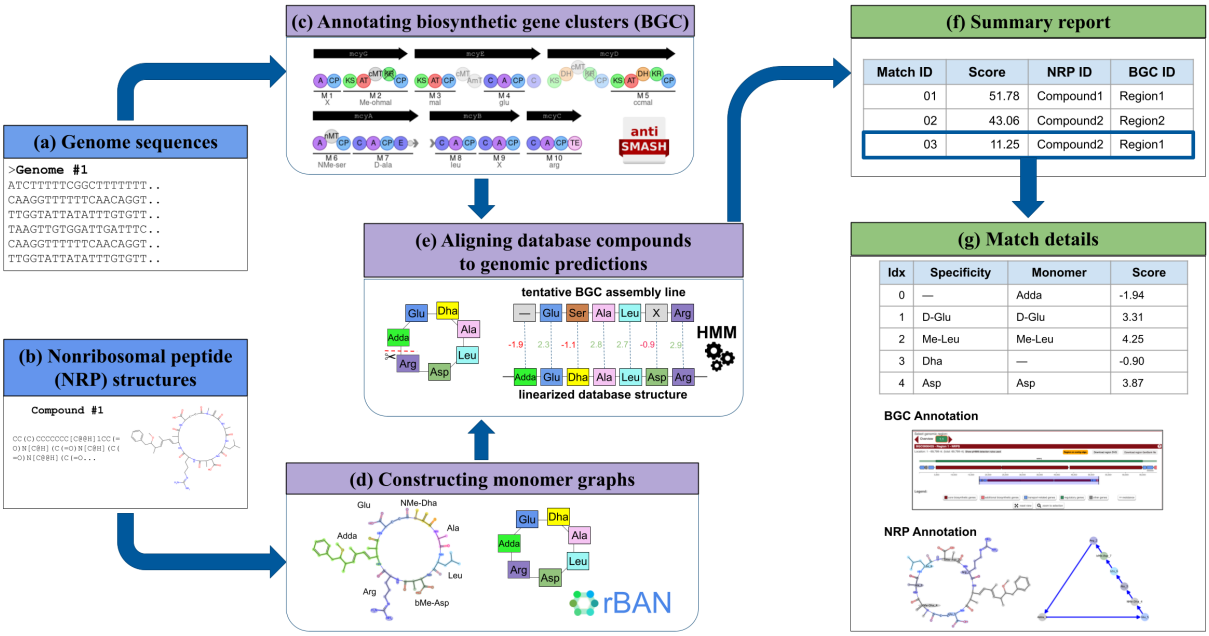

Fig. 1: **Nerpa 2 pipeline**. Input genome sequences (a) and nonribosomal peptide (NRP) structures (b) are preprocessed with third-party tools antiSMASH (c) and rBAN (d) to extract putative biosynthetic gene clusters (BGCs) and NRP monomer graphs, respectively. Nerpa 2 builds HMMs for tentative BGC assembly lines, converts NRP monomer graphs to linear sequences of monomers, aligns them against each other, and scores the alignments (e). The best scoring hits per NRP and BGC are reported (f) with an option to view the detailed information about each match (g).

non-collinear assembly lines and other irregularities of the NRP biosynthesis. BioCAT, introduced later, demonstrated performance comparable with Nerpa and could not solve its critical limitations (Konanov *et al.*, 2022).

Here, we present Nerpa 2, a major upgrade to our previous BGC-NRP linking tool. The new flexible alignment and scoring procedure based on hidden Markov models enabled the handling of complex, non-collinear assembly lines. The Random Forest-based postprocessing of specificity predictions enhanced the tool ability to process promiscuous A domains. We demonstrate how these changes allowed Nerpa 2 to outperform its predecessor in accuracy and sensitivity. Furthermore, Nerpa 2 complements the output with an interactive HTML report, offering users an intuitive and convenient way to interpret and analyze the results.

2. Materials and methods

2.1. Pipeline overview

Nerpa 2 takes as input (meta)genome sequences and NRP chemical structures (Figure 1ab). The data are further handled with the latest versions of the state-of-the-art tools: antiSMASH v.7 (Blin *et al.*, 2023) for the genome mining of BGCs and rBAN v.1 (Ricart *et al.*, 2019) for the retrobiosynthesis of NRP structures (Figure 1cd). Predicted BGCs and disassembled NRP structures are post-processed, converted into linear sequences of NRP building blocks, aligned and scored against each other (Figure 1e). Finally, the tool summarizes the results and reports the best matches (Figure 1fg).

Nerpa 2 operates with NRP building blocks (further referred to as "monomers"), which include amino acids, lipid tails, and other types of residues present in NRP structures. Each monomer is characterized by a core residue and additional features such as its methylation and stereochemistry (D-/L-configuration), when applicable. We manually annotated 249 monomers supported by antiSMASH and 397 monomers from the Norine database used by rBAN to classify them into core residues and additional features. The final monomer dictionary enables streamlined conversion of both preprocessing tools' outputs into the unified set of 42 core residues and a special "unknown" residue.

2.2. Processing of A domains

Given a genome, antiSMASH detects BGCs related to NRP synthesis and annotates them into modules and domains (Figure 1c). For each A domain, antiSMASH extracts 34 amino acid and 10 amino acid signatures, often referred to as the 8Å (Röttig *et al.*, 2011) and Stachelhaus codes (Stachelhaus *et al.*, 1999), respectively. These signatures are then used for the A domain specificity prediction as implemented in the NRPyS module of antiSMASH. First, it uses a set of SVM models from NRPSPredictor2 (Röttig *et al.*, 2011) to predict both a single substrate (single amino acid prediction) and clusters of substrates (small cluster, large cluster, and physicochemical class predictions) from the aa_{34} signature. However, NRPSPredictor2 supports only a limited set of 26 substrates. antiSMASH overcomes this by, second, searching the aa_{10} signatures in the lookup table of 2,319 A domain signatures with known specificity and reporting the closest match.

2.2.1. Raw specificity scores

We grouped the antiSMASH lookup table of signatures [1] by A domain specificity. In doing so, we substituted each substrate with the respective core residue. If a signature had multiple specificities, it was added to each of the relevant specificity groups. This resulted in the lookup table mapping each core residue to a set of associated signatures.

For each A domain, we construct a matrix of features $M(aa_{34}) = [x_1, x_2, \ldots, x_n]$, where x_i is a feature vector associated with the i-th core residue. The feature vector has the structure

$$x_i(aa_{34}) = (s^{(i)}_{single}, s^{(i)}_{small}, s^{(i)}_{large}, s^{(i)}_{class}, s^{(i)}_{10}, s^{(i)}_{34}),$$

where

- $s^{(i)}_{single} \in \mathbb{R}_{>0}$ equals the SVM classification score produced by the "single amino acid prediction" SVM of NRPSPredictor2 if the i-th core residue was predicted as a "single amino acid prediction" or $s^{(i)}_{single} = 0$ otherwise;
- $s^{(i)}_{small}, s^{(i)}_{large}, s^{(i)}_{class} \in \mathbb{R}_{>0}$ equal SVM classification scores produced by the "Small cluster", "Large cluster", and "Physicochemical class" SVMs, respectively, if the i-th core residue belongs to the corresponding cluster or 0 otherwise;
- $s^{(i)}_{10}, s^{(i)}_{34} \in [0, 1]$ are computed from the lookup table as the maximum sequence similarity (the fraction of identical amino acids at corresponding positions) of 10 amino acid or 34 amino acid signatures corresponding to the i-th core residue.

For core residues not supported by the SVMs, we let $s^{(i)}_{single} = s^{(i)}_{small} = s^{(i)}_{large} = s^{(i)}_{class} = -1$.

2.2.2. Aggregate score

We trained a random forest classifier RF that, combined with the lookup table, produces the aggregate score for a feature vector. Let aa_{34} be the 34 amino acid signature of an A domain, and, for each $i \in \{1, \ldots, n\}$, let $\mathcal{T}_i$ be the subset of the 34 amino acid signatures for the i-th core residue in the lookup table, and x_i be the feature vector. Then, the aggregate score y_i is defined as

$$y_i = \begin{cases} 1, & aa_{34} \in \mathcal{T}_i, \\ RF(x_i), & \text{otherwise}, \end{cases}$$

Each score y_i is then calibrated using pretrained parameters from the ground-truth BGC-NRP alignments.

2.2.3. Probability distribution

We model the specificity of the A domain by constructing a probability distribution $\{p_i\}_{i=1,\ldots,n}$ on the set of core residues. This distribution is calculated on the basis of the A domain aggregate scores from the previous section.

Aggregated scores $y_i,\ i \in \{1, \ldots, n\}$ can be viewed as a non-normalized "probability" distribution. Let $Y = \sum_{i=1}^{n} y_i$ be the total "probability". Values of Y greater than 1 indicate that the A domain is likely to be promiscuous, and in this case p_i is obtained by normalizing:

$$p_i = y_i / Y.$$

Values of Y less than 1, on the other hand, mean that the model could not give a reliable prediction for the A domain. In this case p_i is obtained by adding the appropriate fraction of the default distribution:

$$p_i = y_i + (1 - Y) \cdot q_i,$$

where $\{q_i\}_{i=1,\ldots,n}$ is the default distribution of supported core residues computed from the Norine database. A small portion of the default distribution q_i is added to p_i to avoid predictions with zero probability:

$$p_i \leftarrow 0.95 \cdot p_i + 0.05 \cdot q_i.$$

2.3. Construction of tentative BGC assembly lines

Using antiSMASH annotation of a BGC, we classify its genes into *core*—those primarily responsible for producing the BGC's end product(s)—and *non-core* (all others). In non-collinear NRP BGCs, the biosynthetic pathway can deviate substantially from the physical ordering of genes in the genome. Moreover, certain BGCs generate distinct products via multiple assembly lines (see Xu *et al.* (2017)).

To handle these complexities, we employ a systematic procedure for building *tentative assembly lines*, sequences of BGC modules corresponding to plausible orders of gene activation.

1. **Segmentation of Core Genes into Blocks.** We begin by identifying putative start and end points of biosynthesis based on domain-knowledge heuristics. For example, we treat a starter condensation domain as a likely initiation site and a termination thioesterase domain as a likely final step. These markers, however, do not always indicate true biological starts or ends and can appear in the middle of a biosynthetic pathway. We split all the core genes into *blocks*, sets of consecutive core genes bounded by these markers. We treat these blocks as building units for creating tentative assembly lines.
2. **Integration of Gene Blocks into a Complete Assembly Line.** We concatenate the blocks from the previous step into one or more *candidate* continuous assembly pathways. By default, we preserve the genomic order, but we also consider the reverse ordering for blocks on the opposite DNA strand or when domain organization indicates an alternate arrangement.
3. **Extraction and Linearization of Modules with A Domains.** For each block in a candidate pathway, we extract and "chain" the modules that contain A domains in the inferred order, yielding a linear sequence of modules. Each module is further annotated with its predicted specificity, modifying domains (methylation and epimerization), and its relative position on the gene (start, middle, or terminal).

Note that the resulting tentative assembly lines incorporate all core genes containing an A domain and identified by antiSMASH, some of which could be artifacts or belong to neighboring BGCs. To account for this, we retain the information about the borders of gene blocks and have small penalties for skipping entire gene blocks at the ends of a

[1] `https://dl.secondarymetabolites.org/releases/stachelhaus/1.1/signatures.tsv` Accessed 20 Jan 2025.

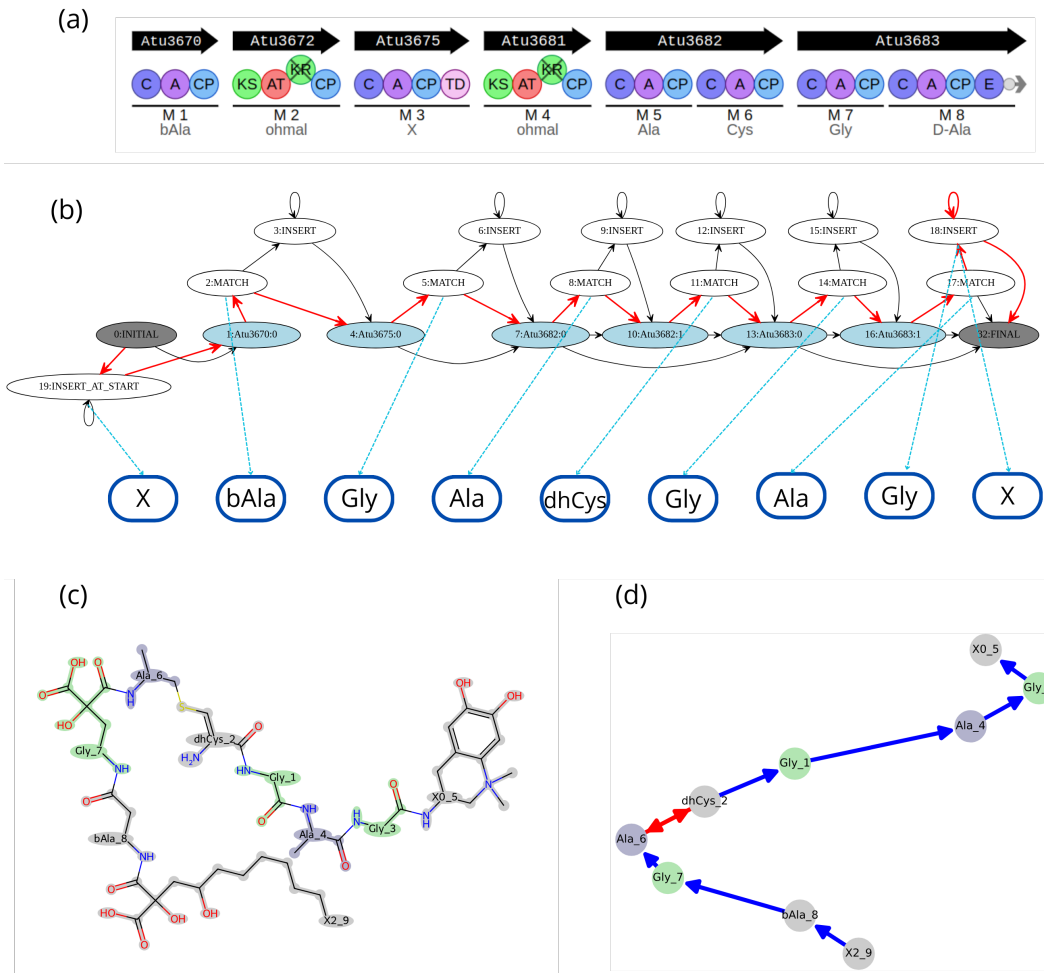

Fig. 2: **Scoring of a biosynthetic gene cluster (BGC) against a nonribosomal peptide (NRP) structure.** The BGC annotation (a) is used to build the hidden Markov model (HMM) for the alignment (b, top). The annotated NRP compound (c) is represented as a monomer graph with blue/red edges depicting peptide/other bonds connecting the monomers (d). The graph is linearized (b, bottom) and scored against the HMM (b). The optimal path through the HMM for the given NRP linearization is highlighted in red. Blue dotted edges indicate HMM emissions. Monomer labels *bAla*, *dhCys*, and *X* denote β-alanine, dehydrocysteine, and unrecognized monomer, respectively. The genes Atu3672 and Atu3681 in the BGC annotation correspond to the non-NRP part of the biosynthetic pathway and are excluded from the HMM. The example is based on the fabrubactin A-BGC0002107 pair from MIBiG (Zdouc *et al.*, 2025; Vinnik *et al.*, 2021).

tentative assembly line in subsequent alignment against NRP linearizations.

This approach allows us to handle complex assembly lines effectively while minimizing the impact of unrelated or misclassified core genes in the BGC.

2.4. HMM construction

For each tentative assembly line, we construct a hidden Markov model (HMM) structured as a linear chain of module-specific subgraphs. Each subgraph corresponds to a module in the assembly line and includes multiple paths that represent different outcomes, such as emitting an NRP monomer, skipping the module, or inserting additional NRP monomers to account for iterative use or non-NRP contributions (Figure 2). The HMM also incorporates additional nodes and edges to handle more nuanced scenarios such as duplicating entire genes or skipping entire gene blocks. Transition weights depend on the genomic context of each module; for example, a module at the end of a fragment has a higher likelihood of additional emissions than one in the middle.

The emission probabilities for each module are calculated as the product of three components, which represent independent selection of the core residue, methylation, and chirality of an NRP monomer:

1. Residue probability, prediction of module's A domain specificity.
2. Methylation probability, the probability that an NRP monomer is methylated conditioned on the presence or absence of a methylation domain in the module.
3. Chirality probability, the probability that an NRP monomer has a specific chirality conditioned on the presence or absence of an epimerization domain in the module.

Emission probabilities for states corresponding to inserting additional NRP monomers do not depend on BGC modules and reflect the independent selection of an NRP monomer's core residue, methylation, and chirality on the basis of the default frequencies in the Norine database.

2.5. Processing of NRP structures

The input chemical structures of NRPs are decomposed into monomer graphs by rBAN (Figure 1d). Then Nerpa 2 cuts the graphs into linear fragments and reassembles them into linear monomer sequences, as follows. First, Nerpa 2 removes all non-peptide bonds, along with any vertices of in- or out-degree zero except leaves (sinks and sources). In practice, this yields no more than four connectivity components, each being a path, a simple cycle, or a path attached to a cycle. For cyclic components, we generate all possible linearizations by choosing different starting monomers. For other shapes, we attempt to find a Hamiltonian path. If no Hamiltonian path is found, the component is discarded; if multiple valid paths exist, we keep only the first. Next, we take every possible permutation of these components and every possible linearization for each component, concatenating them to produce one or more candidate monomer sequences of the original NRP.

2.6. Null hypothesis HMM for an NRP linearization

For each NRP linearization, Nerpa 2 constructs a *null hypothesis* HMM. This HMM comprises a single path of the same length as the linearization. Each state in the path emits an NRP monomer, with the emission probabilities based on the default frequencies, just like in states corresponding to inserting additional NRP monomers in HMMs described in 2.4.

2.7. Scoring HMMs against NRP linearizations

Given an HMM and an NRP linearization, Nerpa 2 calculates the most probable path in HMM that yields the NRP monomers using the Viterbi algorithm (Forney, 1973). Also, Nerpa 2 computes the probability of the NRP monomers being generated by the null hypothesis HMM. The final score is computed as log-odds of the two probabilities.

2.8. Parameters training, manual curation

We selected 80 BGCs from the MIBiG v.4.0 database labeled as producing NRPs, for which the referenced literature provided tentative assembly lines. From these entries, we manually curated 127 "ground-truth" alignments between BGCs and their NRP products, since some BGCs can produce multiple compounds.

These ground-truth alignments were used to calibrate the A domain specificity prediction model using empirical frequencies. Specifically, we evaluated how often the predicted amino acid incorporations matched the ground truth alignments for different score ranges and used these empirical frequencies to adjust the prediction model, thereby improving its reliability.

The ground truth alignments were also used to set the weights of HMM transitions and emission probabilities. Probabilities of BGC module skips and insertions of extra NRP monomers, as well as the likelihood of methylation and D chirality, given the presence or absence of methylation and epimerization domains, were calculated based on the respective empirical frequencies.

2.9. Reporting the results

Nerpa 2 presents the results in machine-readable JSON files, suitable for the tool integration into large computational pipelines, and in a user-friendly HTML report, suitable for manual analysis of the results (Figure 1fg; Supplementary Figure S1). The HTML report offers a summary table of Nerpa 2 matches, which enables sorting and filtering them by score, BGC, or NRP (Figure 1f). The summary table is complemented with a detailed view featuring the alignment information, antiSMASH output for the corresponding BGC, and depiction of the matched NRP as an annotated molecular structure and a monomer graph (Figure 1g). The HTML report is built with JavaScript, and the molecular structure and monomer graph figures are drawn using RDKit (Landrum *et al.*, 2006) and NetworkX (Hagberg *et al.*, 2008) Python libraries.

3. Results

3.1. Benchmarking dataset

The target NRP chemical database was formed from Norine structures (Flissi *et al.* (2020); accessed 19.11.2024) and MIBiG v.4.0 annotations of NRP-related BGCs. This set was preprocessed with rBAN v.1 (Ricart *et al.*, 2019), and structures with less than three annotated monomers were excluded. For the sake of speed, the database was further deduplicated based on the monomer graph isomorphism on the set of Nerpa-supported monomers. This procedure resulted in 1,205 unique NRP chemical structures.

To form the genomic side of the benchmark dataset, we downloaded all 877 NRP-related entries from MIBiG v.4.0 (Zdouc *et al.*, 2025) in the GenBank format and processed them with antiSMASH v.7.1.0 for consistency. We further narrowed this set down to 296 BGCs having at least three A domains and the corresponding chemical structures with at least three monomers annotated by rBAN. These BGCs were clustered using BiG-SCAPE v.1.1.9 (Navarro-Muñoz *et al.*, 2020) and all BGCs that form a gene cluster family with BGCs from our training set (Section 2.8) were removed. The resulting benchmarking dataset consists of 191 BGCs.

3.2. Accuracy of BGC-NRP linking

We compared Nerpa 2 performance against Nerpa 1, which was adopted to work with antiSMASH v.7 to exclude the potential improvements in the third-party genome mining step and thus make the core algorithms comparison fair. We excluded BioCAT from the benchmarking since this tool was prohibitively slow on our test dataset due to a lack of parallelization, and this tool can work only with antiSMASH v.6.

Every of the 191 benchmarking BGCs was compared against the entire chemical structure database, and the 10 best-scoring matches per BGC were retained for each tool. The match was considered correct if the linked NRP was among the MIBiG ground truth annotations for the corresponding BGC and the rank of the correct match was recorded.

Figure 3 demonstrates the performance of the tools from different perspectives. Nerpa 2 is consistently better than the predecessor in linking BGCs to their correct NRP annotations with 50.3% accuracy at rank=1 that gradually increases to 77.0% at rank=10 compared to 41.9% and 67.0% for Nerpa 1, respectively (Figure 3a). The receiver operating characteristic curve shows that both tools are on par in scoring correct matches

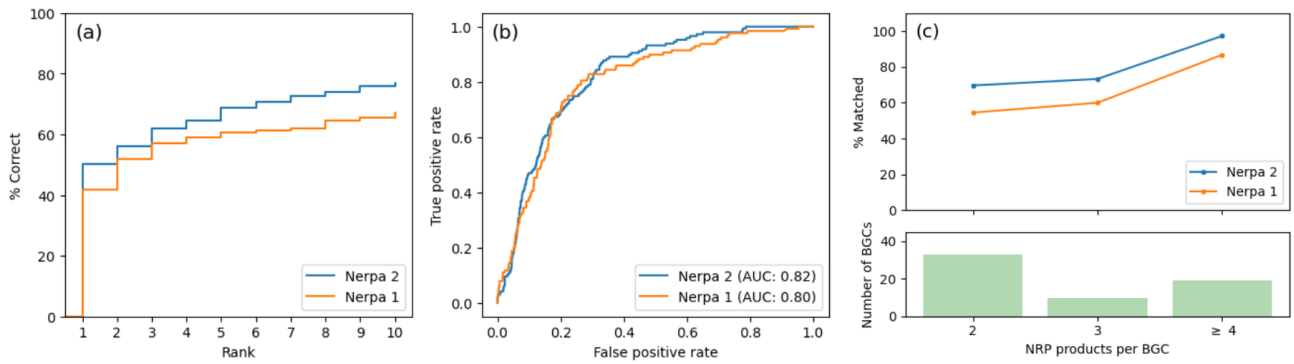

Fig. 3: **Nerpa 1 (orange) and 2 (blue) accuracy in the benchmark experiment.** (a) percentage of correct identifications in the top n matches, (b) receiver operating characteristic curves, and (c) mean percentage of matched NRP products for the BGCs producing multiple compounds.

better than false positive identifications (Figure 3a). Finally, Nerpa 2 outperforms the previous version in annotating BGCs linked to multiple NRP products, a common scenario for BGCs containing promiscuous A domains (Figure 3c).

3.3. Inspecting performance on non-standard assembly lines

Improvements in the alignment algorithm allow Nerpa 2 to handle complex non-collinear NRP assembly lines. Supplementary Figures S2 and S3 illustrate the alignments produced by Nerpa 1 and Nerpa 2 for an interactive NRP (fuscachelin B, BGC0000359 (Dimise *et al.*, 2008)) and an assembly line missing some modules and reusing one of the genes (didemnin Y, BGC0000985 (Xu *et al.*, 2012)). While Nerpa 1 fails to match most NRP monomers properly in these cases, the new version produced the correct alignments in both.

4. Conclusion

We have introduced Nerpa 2, a tool for high-throughput linking nonribosomal peptide (NRP) biosynthetic gene clusters (BGCs) to their putative products. By integrating a probabilistic model for predicting Adenylation domain specificities and utilizing a flexible HMM-based alignment scoring scheme, Nerpa 2 effectively addresses the challenges posed by the NRP biosynthesis, including domain promiscuity and non-collinear assembly lines.

Our benchmarking experiments indicate that Nerpa 2 outperforms its predecessor in both accuracy and sensitivity, particularly in handling complex cases that were previously unsupported. Beyond these enhanced capabilities, Nerpa 2 offers a user-friendly interface, generates easy-to-interpret output, and is fully cross-platform compatible. The tool accepts (meta)genome sequences in FASTA and GenBank and supports direct import of antiSMASH website results. Users can provide an unlimited number of NRP structures in the SMILES and InChi formats or use our pre-processed database of known and tentative NRP structures.

We anticipate that Nerpa 2 will become a valuable utility for the NRP research community, aiding in the annotation of computationally predicted BGCs from isolated organisms and microbiome samples. Additionally, Nerpa 2 can serve as an effective dereplication tool for filtering out BGCs likely linked to known compounds and paving the way for the discovery of novel NRPs.

Supplementary data

Supplementary data are available online.

Conflict of interest

None declared.

Funding

Saarland University [the NextAID project to support the position of A.T.].

References

Agrawal, S., Acharya, D., Adholeya, A., Barrow, C. J., and Deshmukh, S. K. (2017). Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential. *Frontiers in pharmacology*, **8**, 828.

Bağcı, C., Nuhamunada, M., Goyat, H., Ladanyi, C., Sehnal, L., Blin, K., Kautsar, S. A., Tagirdzhanov, A., Gurevich, A., Mantri, S., *et al.* (2024). Bgc atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. *Nucleic Acids Research*, page gkae953.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019). antismash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research*, **47**(W1), W81–W87.

Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J., *et al.* (2023). antismash 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic acids research*, **51**(W1), W46–W50.

Dejong, C. A., Chen, G. M., Li, H., Johnston, C. W., Edwards, M. R., Rees, P. N., Skinnider, M. A., Webster, A. L., and Magarvey, N. A. (2016). Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol*, **12**(12), 1007–1014.

Dimise, E. J., Widboom, P. F., and Bruner, S. D. (2008). Structure elucidation and biosynthesis of fuscachelins, peptide siderophores from the moderate thermophile thermobifida

fusca. *Proceedings of the National Academy of Sciences*, **105**(40), 15311–15316.

Flissi, A., Ricart, E., Campart, C., Chevalier, M., Dufresne, Y., Michalik, J., Jacques, P., Flahaut, C., Lisacek, F., Leclère, V., *et al.* (2020). Norine: update of the nonribosomal peptide resource. *Nucleic acids research*, **48**(D1), D465–D469.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, **61**(3), 268–278.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, Š., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in streptomyces ambofaciens. *Chemistry & biology*, **16**(4), 421–431.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J., Van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., *et al.* (2020). Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research*, **48**(D1), D454–D458.

Konanov, D. N., Krivonos, D. V., Ilina, E. N., and Babenko, V. V. (2022). Biocat: Search for biosynthetic gene clusters producing nonribosomal peptides with known structure. *Computational and Structural Biotechnology Journal*, **20**, 1218–1226.

Kunyavskaya, O., Tagirdzhanov, A. M., Caraballo-Rodríguez, A. M., Nothias, L.-F., Dorrestein, P. C., Korobeynikov, A., Mohimani, H., and Gurevich, A. (2021). Nerpa: a tool for discovering biosynthetic gene clusters of bacterial nonribosomal peptides. *Metabolites*, **11**(10), 693.

Landrum, G. *et al.* (2006). Rdkit: Open-source cheminformatics.

Medema, M. H., de Rond, T., and Moore, B. S. (2021). Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics*, **22**(9), 553–571.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L., Yeong, M., Cruz-Morales, P., Abubucker, S., *et al.* (2020). A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology*, **16**(1), 60–68.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3), 443–453.

Ricart, E., Leclère, V., Flissi, A., Mueller, M., Pupin, M., and Lisacek, F. (2019). rban: retro-biosynthetic analysis of nonribosomal peptides. *Journal of cheminformatics*, **11**(1), 1–14.

Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). Nrpspredictor2—a web server for predicting nrps adenylation domain specificity. *Nucleic acids research*, **39**(suppl_2), W362–W367.

Sieber, S. A. and Marahiel, M. A. (2003). Learning from nature's drug factories: Nonribosomal synthesis of macrocyclic peptides. *Journal of Bacteriology*, **185**(24), 7036–7043.

Stachelhaus, T., Mootz, H. D., and Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol*, **6**(8), 493–505.

Süssmuth, R. D. and Mainz, A. (2017). Nonribosomal peptide synthesis—principles and prospects. *Angewandte Chemie International Edition*, **56**(14), 3770–3821.

Vinnik, V., Zhang, F., Park, H., Cook, T. B., Throckmorton, K., Pfleger, B. F., Bugni, T. S., and Thomas, M. G. (2021). Structural and biosynthetic analysis of the fabrubactins, unusual siderophores from agrobacterium fabrum strain c58. *ACS Chemical Biology*, **16**(1), 125–135.

Xu, F., Nazari, B., Moon, K., Bushin, L. B., and Seyedsayamdost, M. R. (2017). Discovery of a cryptic antifungal compound from streptomyces albus j1074 using high-throughput elicitor screens. *Journal of the American Chemical Society*, **139**(27), 9203–9212.

Xu, Y., Kersten, R. D., Nam, S.-J., Lu, L., Al-Suwailem, A. M., Zheng, H., Fenical, W., Dorrestein, P. C., Moore, B. S., and Qian, P.-Y. (2012). Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *Journal of the American Chemical Society*, **134**(20), 8625–8632.

Zdouc, M. M., Blin, K., Louwen, N. L., Navarro, J., Loureiro, C., Bader, C. D., Bailey, C. B., Barra, L., Booth, T. J., Bozhüyük, K. A., *et al.* (2025). Mibig 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic acids research*, **53**(D1), D678–D690.