Creating a Natural Product Database and Generating Molecular Substructures

Sam Stokman
MSc Thesis Report
25-01-`19
Supervisors:
dr. Justin J.J. van der Hooft
dr. Marnix Medema

Abstract

Natural products (NPs) are an important source of therapeutic agents whose innovation and improvement depends on NP discovery. The knowledge about the assembly of NPs increases as biosynthetic gene clusters (BGCs) are identified and annotated. However, the majority of BGCs have yet to be linked to the compounds they encode for. To study unknown NPs, the analytical method of choice is often tandem mass spectrometry (MS/MS). The disadvantage of using MS/MS for NP discovery is the lack of reference spectra which makes interpreting newly analyzed structures very challenging. The BGC and mass spectra interpretation problems are a big bottleneck for structural NP elucidation. A promising solution to enhance successful NP elucidation is the recognition and annotation of NP biochemical building blocks (substructures). Ideally, the chemical composition of the recognized substructure is identical to the precursor in BGCs and the annotated mass fragment in MS data as defined by others. In order to generate all relevant substructures, it is important to cover the largest number of NP structures possible and therefore, first NPDatabase was built that stores more than unique 320,000 NPs. The most well-known fragmentation methods, using a top-down approach (i.e. BRICS and RECAP), result substructures without similar chemical compositions to those found in BGCs and MS data. To generate higher quality substructures, a new fragmentation ruleset was developed, NPRules. NPRules is implementable in the structure fragmentation program molBLOCKS and highly suitable for NPs from saccharide, ester and peptide classes. Additionally, an alternative and opposite (bottom-up) approach for structure fragmentation is proposed. This method first generates pre-substructures which then are filtered through several steps such that only the qualitative substructures remain. The bottom-up method is not optimal yet but does offer a new perspective and can, after optimization, be used if the top-down method fails. NPRules creates more relevant substructures than the other top-down methods confirmed by validation with identified and annotated BGCs and mass spectra. NPRules results in equal substructure quantities but the substructure quality is better as the recall and precision are higher. Generation of relevant substructures for all NPs in NPDatabase allows linkage from precursors in BGCs and mass fragments in mass spectra to the candidate structure they originally represented. The recognition and annotation of NP substructures as described here enhances successful NP elucidation and tackles a big problem in the genomics and metabolomics fields.

Contents

1. Introduction

Natural products (NPs) are specialized molecules produced by living organisms. NPs are synthesized by microorganisms, for example in the mammalian gut, in and on plant roots and in the marine environment. A broad range of scientists, in fields including drug discovery, ecology, biosynthesis, and chemical biology, show a major interest in NPs as they are used in the commercial product development for human medicine, animal health, and plant crop protection [1]. Between 20–25% of all approved therapeutic agents today trace their origins to NPs by providing or inspiring the development. These numbers are even higher in the field of cancer where they reach 75%. For the innovation and improvement of human medicine and other commercial products, NP discovery and therefore, NP research will always be of great importance [2].

The knowledge about the assembly of NP compounds increases, together with the growing accessibility of genome sequence data. Biosynthetic gene clusters (BGCs) are identified by the recognition of physically clustered genes that encode for the enzymes responsible for NP assembly [3]. However, a majority of BGCs remain orphan, as they have yet to be linked to the compounds they encode for [4]. Linkage of NP compounds to BGCs enables genome sequence data to facilitate NP discovery.

To study unknown metabolic substances often untargeted analysis is used which takes all detected metabolite features into consideration. The advantage of this high-throughput approach is that no prior knowledge about the structure's identity is required. Tandem mass spectrometry (MS/MS) combined with liquid chromatography is often the analytical method of choice; the (unknown) analyzed chemical structures are represented by mass/ fragmentation spectra [5]. The disadvantage of using MS/MS for NP discovery is the lack of reference spectra which makes interpreting newly analyzed structures very challenging. Recently, a study has introduced an infrastructure to enable sharing and curation of raw, processed and identified MS data which stimulates knowledge sharing within the NP community [6]. Another recent study, focusing on interpreting mass spectra, has introduced a metabolome mining strategy that extracts biochemically relevant molecular substructures ("Mass2Motifs") as sets of co-occurring mass fragments and neutral losses in an unsupervised manner [7]. NP studies as such are a step in the right direction, however, interpreting mass spectra remains a big bottleneck for structural NP elucidation which leads to unanswered questions about NP diversity and identity till today.

A promising solution to enhance successful NP elucidation is the recognition and annotation of NP biochemical building blocks (substructures). These substructures play a key role in linking BGCs and mass spectra to the NP structures they represent. Ideally, the chemical composition of the substructure is identical to the precursor in BGCs and the annotated mass fragment in MS data as defined by others [8-29].

Several open source tools that provide substructure generation use fragmentation rules. These rules specify bond breakage between atoms or atom groups defined by Smiles Arbitrary Target Specification (SMARTS) patterns. The best-known fragmentation methods are breaking retrosynthetically interesting chemical substructures (BRICS) and retrosynthetic combinatorial analysis procedure (RECAP). BRICS creates more fragments than RECAP and also the number of fragments with multiple connection point is higher [30]. The latter is considered as a disadvantage since ring structures are less preserved and the number of undesired substructures increases. RECAP does preserve ring structures but is not comprehensive enough as crucial bonds are missed, also undesirable side chains cleavage occurs. Another well-known rule in the field is the CCQ rule. This single rule cleaves the bond between two carbons of which one is next to a hetero atom. The resulting substructure quantities are sufficient but the specificity lacks as this rule is too general [31]. These different rulesets are implemented in the suite molBLOCKS [32]. molBLOCKS was optimized by Heikamp et. al as new parameters were implemented and a new ruleset, extendedRECAP (redefined and extended RECAP rules), was added. ExtendedRECAP is more comprehensive than RECAP but also increases undesirable side chain cleavage.

For this project, first the open-source NPDatabase was build that stores more than 320,000 unique NP structures. NPDatabase can be used by others and was the starting point for this project, as the structures in it were fragmented for the generation of substructures. Since the CCQ, BRICS, RECAP and extendedRECAP rulesets did results substructures without similar chemical compositions to those found in BGCs and MS data, a new set of fragmentation rules was developed, NPRules. NPRules is a more specific ruleset, highly suitable for NPs from saccharide, ester and peptide classes, that is easy to use as it is implementable in the molBLOCKS suite. NPRules performs better than the other rulesets confirmed by validation with identified and annotated BGCs and mass spectra. Additionally, an alternative and opposite (bottom-up) approach for structure fragmentation is proposed. This method does not depend on external tools or fragmentation rules and is applicable for a wide range of NP structures. This bottom-up approach is not optimal yet, because of the high computational burden, but it definitely deserves thought and deeper study.

2. Methods and Implementation

Data sources and processing

The NP data from eleven external databases was collected which resulted in a total of 497,610 structures (figure 1). The data included overlapping structures within and between databases. The minimum input data required was the structure's simplified molecular-input line-entry specification (SMILES) and its external database identifier. Three open source tools provided NP structure processing in Python programming language. The RDkit library (www.rdkit.org) was used to generate the SMILES, International Chemical Identifier (InChI), molecular weight and molecular formula. Based on the canonical SMILES, the number of unique input structures was reduced to 322,242. The standardized InChIKeys were generated by the JChem tool from ChemAxon (www.chemaxon.com, in collaboration with Rutger Ozinga). The structures were classified according to kingdom, superclass, class and subclass by ClassyFire (classyfire.wishartlab.com, in collaboration with Oscar Hoekstra).

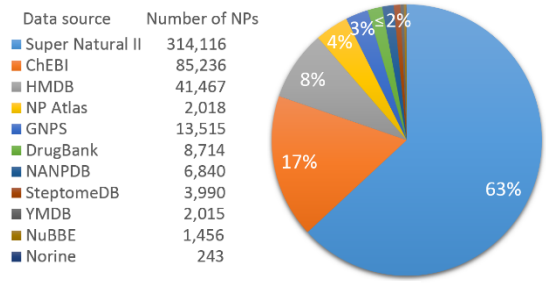


Figure 1. NP data sources.

```
     structure_id = NP_16830
           smiles = C=C1CCC(C)(C2=CCC(CO)=CC2)C1C
molecular_formula = C15H22O
monoisotopic_mass = 218.16707
           inchi = InChI=1S/C15H22O/c1-11-8-9-15(3,12(1
       inchi_key = FFSGCOHFSZTTRC-SWLSCSKDSA-N
         kingdom = Organic compounds
      superclass = Organic oxygen compounds
           class = Organooxygen compounds
        subclass = Alcohols and polyols
```

Figure 2. Example structure stored in the Structure table showing the attributes and data.

Database design and content

All generated data from 322,242 NP structures was implemented in NPDatabase (figure 3) with SQLite in python programming language. NPDatabase can be downloaded from www.dropbox.com/s/qumnikhiaszrwjh/NPDatabase.sqlite?dl=0 and contains six tables. Since NPDatabase is kept up to date, small alterations between versions can occur. The data in the database can be extracted with Structured Query Language (SQL). The Structure table stores the generated data mentioned above together with its primary key, the structure_id. The Structure_has_data_source and the Data_source tables provide external source information. The attributes source_name and source_id enable the user to easily search for the structure and its original data source online. Data of an example NP structure stored in the Structure table is shown in figure 2. The relevant substructures are not yet included, however, NPDatabase is prepared for substructure storage as the Structure_has_substructure and Substructure tables are designed to store all substructural information. When relevant substructures represented with their SMILES are created and collected with the fragmentation methods explained later on, all other Substructure attributes can be generated correspondingly to the Structure attributes. The attribute nr_of_matches in the junction table is the number of substructure matches within the structure. Ultimately, NPDatabase will enable substructure mapping onto candidate structures.

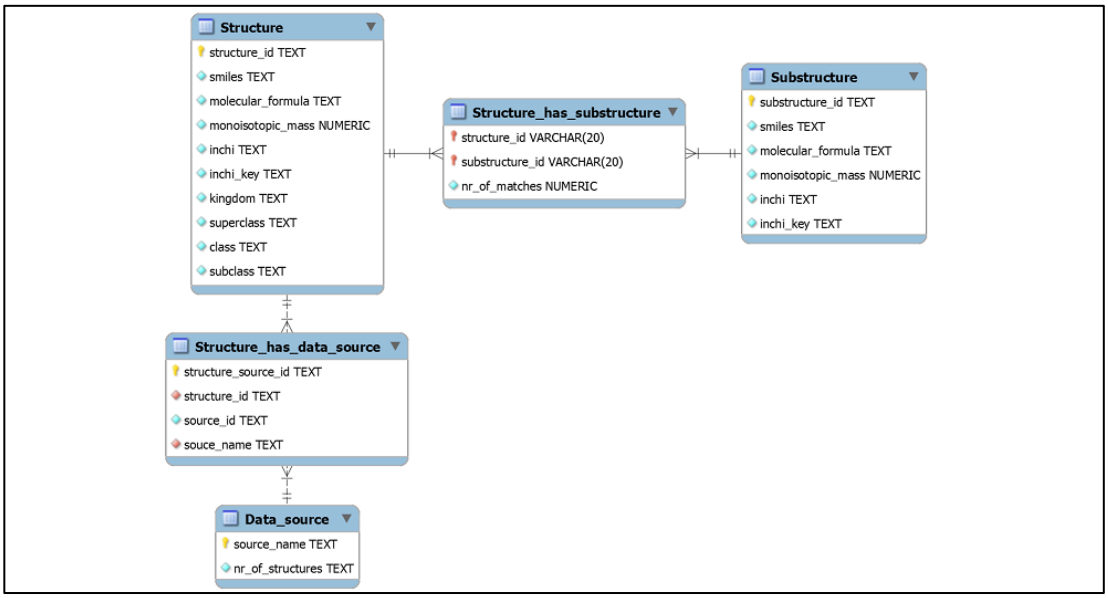


Figure 3. SQL schema for NPDatabase. The structure table contains all NP data including the canonical SMILES. The structure table is linked to the data_source table through the Structure_has_data_source table which provide information about the external data source where the structures originate from. All relevant NP substructure data can be stored in the Substructure table.

Generating substructures

To fragment the NP structures present in NPDatabase into substructures two methods were implemented. The first method makes use of a top-down approach and the second method makes use of a bottom-up approach. The structure's input for both methods is the canonical SMILES and its structure identifier. The workflow schemes for both methods are shown in figure 4, the methods are explained in more details in the sections below.

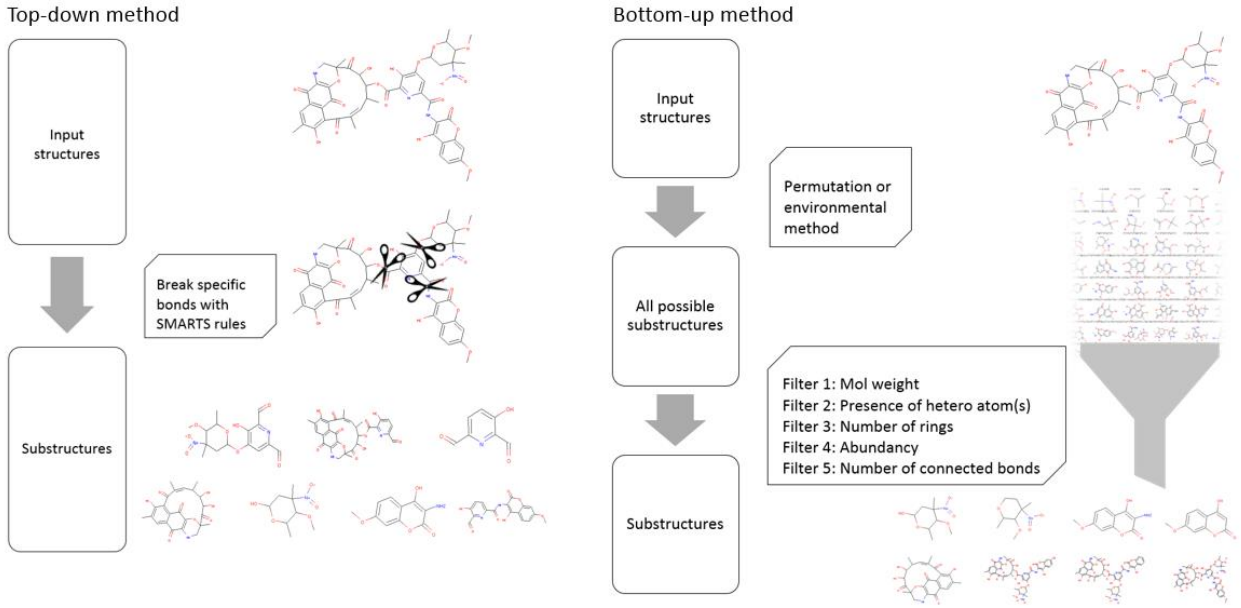


Figure 4. Workflow schemes for the top-down and bottom-up structure fragmentation methods.

Top-down approach

The top-down approach is a method that directly fragments the structure into one or more substructure(s). The user-friendliness of several tools working with this approach, e.g. eMolFrag using BRICS [33] and the RECAP and BRICS Implementation from RDkit were tested. Ultimately, the most user-friendly tool tested was the modified version of molBLOCKS [31]. MolBLOCKS was preferred among other tools because of the possibility to add and test new fragmentation rules, the ability to easily adjust fragmentation parameters, and the convenient output format (the substructures are represented by SMILES). The four default molBLOCKS rulesets (CCQ, BRICS, RECAP and extendedRECAP) did not perform sufficient since they resulted substructures without similar chemical compositions to those found in BGCs and MS data. Another challenge faced was the wide variety of NP structures, it is almost impossible to expect that one ruleset results perfect substructures for each NP type (i.e. proteins, lipids and carbohydrates). For that reason a more specific approach was chosen and a new fragmentation ruleset was defined, called NPRrules. NPRules contains 45 rules that mainly focus on structures with cyclic compounds. To optimally preserve the ring structures, only cyclic compounds that are connected through one bond are targeted. The sidechains are preserved. Acyclic substructures are not fully ignored as pentane and pentene (and longer carbon chains) are also taken into account. NPRules contain 9 rules that target glycosidic bonds, 14 rules that target ester bonds and 2 rules that target amide bonds. For that reason, NPRules is highly suitable for NP structures from saccharide, ester and peptide classes. The other rules are aromatic or cyclic and carbon and/or nitrogen and/or oxygen related, which are abundant in NPs. The balance between fragmenting the input structure too little or too much was found by visual substructure examination after fragmentation with the addition of each rule. The new fragmentation rules, that together represent the NPRules ruleset, are listen in SI Appendix, table S-1.

Since the molBLOCKS suite was chosen for substructure generation and NPRules implementation, several standard parameters values had to be selected. The parameters that determines the maximum molecular weight (g/mol) of a substructure was set to 1000 (-w), and the number of fragments that should be connected and considered as new fragment was set to 1 or 2 (-k). The minimum number of atoms in the substructure was set to 5 (-n), the maximum was set to 100 (-m) and the fragment size relative to the parent structure was set to 0.99 (-s). The exact parameter settings can be found in the result section if other values than the standards mentioned here were used.

Bottom-up approach

To offer an alternative fragmentation method, which is less specific and does not rely on external tools or pre-defined fragmentation rules, a bottom-up approach to create fragments was developed. The bottom-up approach first generates a wide range of pre-substructures and then filters them through five steps such that only the qualitative substructures remain. The RDkit library is required for this approach and two different methods; the permutation method and the environmental method, are implemented for the pre-substructure generation step.

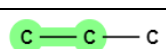
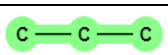
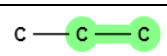
Permutation method

The permutation method actually generates all possible substructures based on the SMILES string. First all combination from two characters on until the full length of the SMILES string (parentheses and bond characters included) are gathered. A SMILES string with a length of 6 characters has 56 combinations. Only valid SMILES strings are tested whether they match the original structure or not. All substructures that indeed do match are considered pre-substructures. The permutation method is recommended for SMILES with a maximum length of 18 characters since the execution time increases exponentially (SI Appendix, section S1.2).

Environmental method

The second method to generate pre-substructures, and recommended for longer input SMILES, is the environmental method. This method first creates substructure environments based on atom number and radius. Each atom in a structure has been assigned an atom number and the radius is defined as the number of atoms that should be included in that specific radius. The total number of atoms in a structure is used to create all atom number and radius combinations that determine the environment. Due to the radius some substructures are not taken into consideration. A simple example is given in table 1 for the structure propane that consist of 3 atoms (hydrogens excluded). Therefore, the atom numbers are 0, 1 and 2 (from left to right considering the SMILES 'CCC'), which also are used for the values of the radii. All possible atom-radius combinations together with the resulting environments (highlighted in green) and substructure are shown in the table below.

Table 1. Example for the structure 'CCC' for generating substructures with the environmental method.

Atom-radius combination	Environment	Substructure
0-0; 1-0; 2-0		-
0-1		CC
0-2; 1-1; 1-2; 2-2		CCC
2-1		CC

Ultimately, these environments are used to generate the corresponding substructures. Again only valid SMILES strings are considered as pre-substructures. The environment method has a shorter execution time than the permutation method (0.13 seconds for a SMILES string of 20 characters versus 3.00 seconds).

Substructure filtering

The pre-substructures then are filtered through several steps. The parameter settings in all steps can be adjusted. The first filter step eliminates substructures which are too small and checks if the structure has a certain molecular weight. The second filter step checks whether the substructure contains hetero atoms. This step is included since hetero atoms (especially halogens) are more distinctive in MS analysis than hydrogen or carbon atoms. In step 3, the number of rings in the substructure is determined, which can be beneficial for cyclic structures. The fourth step checks the abundancy of the substructure. The abundancy here is defined to be the percentage of input structures that contains the substructure. This to eliminate too general or too specific substructures. Note that this filter step totally depends on the input structures as a total. The last filter step checks the substructure's number of bonds which are connected to the rest of the structure. This step is important regarding MS analysis. In general, a substructure with little connected bonds (≤3) is more likely to appear as actual mass fragment than substructures having more connected bonds.

The bottom-up approach for structure fragmentation was used with plausible and consistent but not necessary optimal filter parameters settings (molecular weight ≥40 g/mol, hetero atoms ≥ 0, number of rings ≥ 1, abundancy 0-100, number of connected bond ≤ 2), this to provide comparison possibilities between different NP classes. The python scripts for generating (both methods) and filtering substructures can be downloaded from https://github.com/SamStokman/NPThesis.

Method validation

Substructure quantity

Substructure quantity analysis was conducted first in order to determine the ability of each fragmentation method to fragment NP structures. The NP input structures were selected from NPDatabase based on their class. NP classes contain a sufficient number of input NP structures as well as information about the structure characteristics. Using different NP classes allowed comparison between fragmentation methods and assessment of the performance of each method for each class.

Substructure quality

The quality of the resulting substructures generated with the top-down method, in order to validate NPRules, was manually examined and compared to 'correct' substructures. These correct substructures were biosynthetic precursors and annotated mass fragments defined by others [8-29]. The two different information sources were kept separated and the results of the other top-down fragmentation rules were also taken into account. Since NPRules was created with a purpose to be used for more specific NP classes, mainly NPRules suitable validation structures were used.

For BGC based validation, 12 NP structures [8-18] (SI Appendix, section S1.3, figure S-3 – S-14) were fragmented with the top-down approach using all fragmentation rulesets (NPRules, CCQ, BRICS, RECAP and extendedRECAP). The molBLOCKS parameters values were set as mentioned before. The resulting substructures were regarded as visually correct/ recognized if they were identical to the validation substructures. To allow small substructure alterations also non-identical substructure were examined. These non-identical substructures were still regarded as correct if four condition were met. The first condition was that the substructure can differ one (more or less) atom at the connection point between the substructure and the structure. The second one states that cyclic structures must be closed and thus no bonds should be missing. The third condition required that side chains must be included and the last one states that double substructures within the same structure count as one correct substructure.

Next to BGC based validation, also MS data was used for validation purposes. 12 NP structures [19-29] (SI Appendix, section S1.4, figure S-15 – S-26) were fragmented with the top-down method, again using the five different fragmentation rulesets. The correctness of a generated substructure was based on the correlation between the mass to charge ratio (m/z) of the mass fragment found in the reference mass spectra and the exact mass of that resulting substructure, also neutral losses were taken into account. In theory, the m/z in a mass spectra and exact mass of the same substructure do not have exactly the same value, this is due to the differences caused by the cation fragment in MS analysis (often minus 1.008 for a missing hydrogen atom), variation in MS analysis methods, and decimal rounding. Also, if the mass does match, there is no guarantee that it represents the actual substructure, as it can have the same chemical elements and number of atoms but differs in structural formula. For this reason the substructures were also examined whether they were visually correct/ recognized using the same conditions as set for the BGC validation method.

3. Results

Quantitative fragmentation results

The top-down and bottom-up methods were both used to fragment the structures from the flavonoid, polypeptide, polysaccharide and lactone classes. For the bottom-up method only the environmental approach was used due to the length of the SMILES strings. The results for all four classes are shown in figure 5. The values above the bars represent the generated number of unique substructures.

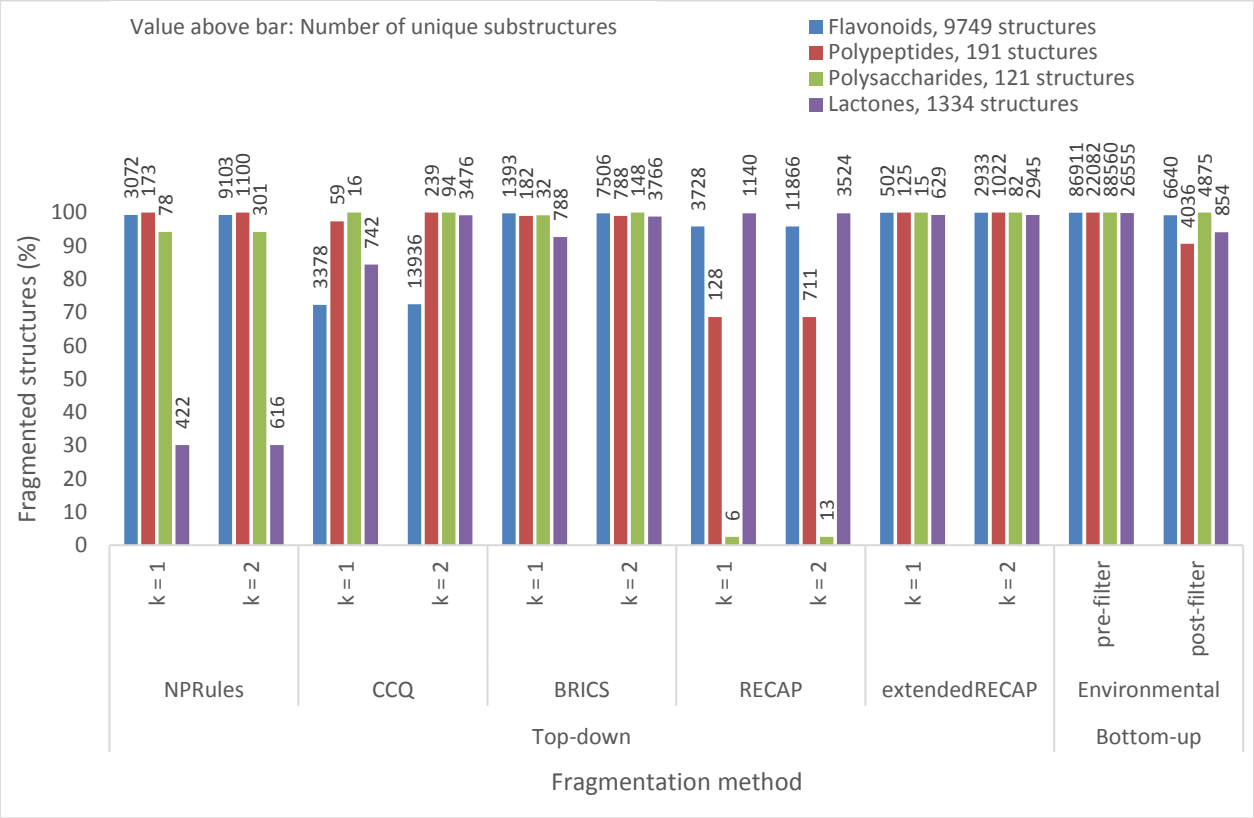


Figure 5. Quantity results flavonoid, polypeptide, polysaccharide and lactone class NP structures.

For the flavonoid, polypeptide and polysaccharide classes, the number of fragmented structures and unique substructures generated using NPRules is comparable to those of the other fragmentation methods. RECAP performs poor for the polysaccharide class. Taking only the top-down methods into consideration, extendedRECAP results the highest (sub)structure quantities followed by BRICS. RECAP generated the lowest number of substructures what was to be expected according to literature. Looking at the results of each ruleset individually, k=1 and k=2 have similar values for the number of fragmented structures. The number of unique substructure generated increases for each ruleset if k=2 was used. As to be expected, the lactone results obtained using NPRules show fewer fragmented structures compared to the other methods, this is due to the complex cyclic structures in lactones as they are connected through several bonds instead of just one. This is an example of a NP class where the bottom-up method could be used instead. Without filtering, this method fragments 99-100% of the input structures and generates the highest number of unique substructures for each class. These substructures were filtered (post-filter) with example parameter settings which show that this method has the capability to perform fragmentation with similar (sub)structures quantity results as the classical top-down methods.

To assess the quantitative results of the bottom-up method using the permutation approach to generate pre-substructures, non-class specific structures (2025 in total) with SMILES strings with a maximum length of 10 characters were fragmented. To make a fair comparison between the bottom-up and the top-down method results, the top-down parameter that determines the minimal number of the substructure's atoms (-n) was also set to 2 (figure 6).

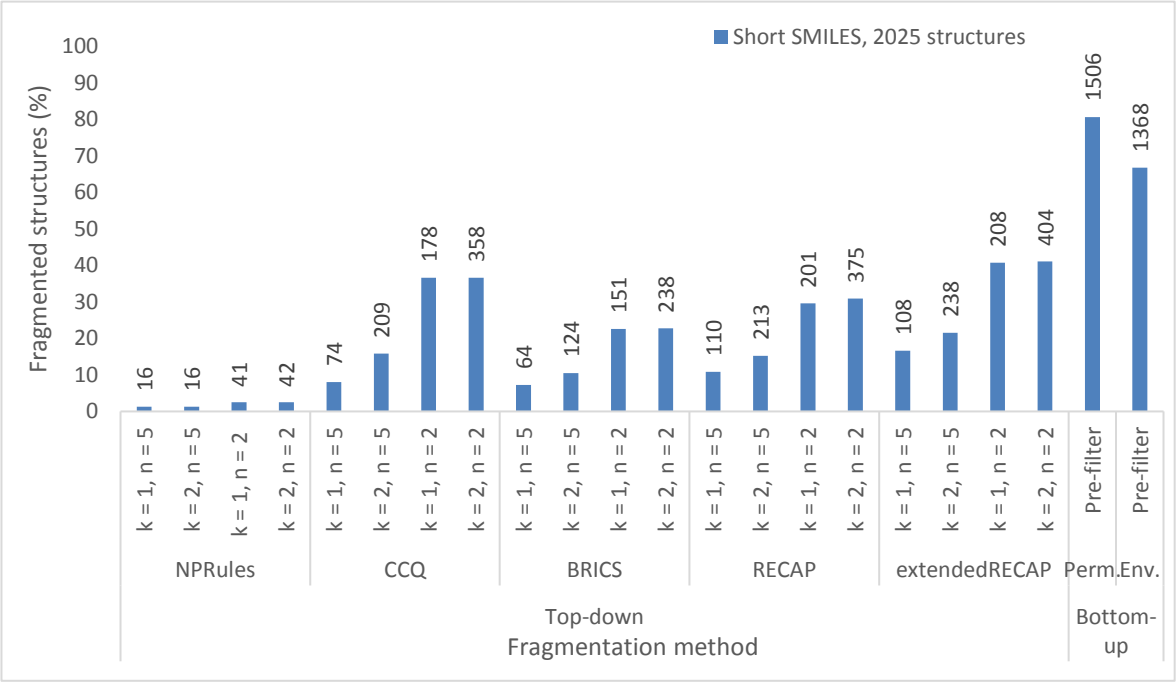


Figure 6. Quantity results for NP structures that have short SMILES strings (length ≤10).

The results show that the permutation method fragments the most structures and generates the highest number of unique substructures followed by the environmental method. These findings do meet the expectations and prove that the bottom-up method can certainly be useful. Note that the substructures are not filtered, this is to show the maximum capacity for quantitative substructure generating. The permutation bottom-up method does not fragment 100% of structures because (sub)structures with a length of one or two characters are excluded.

BGC based validation results

The 12 reference structures used are provided in SI Appendix, section S1.3. Because of the extensive amount of data, the results of one example structure for all top-down methods with k =1 are showed in figure 7. The substructures are represented with a green or red colour. The green substructures in figure 7a are regarded as 'correct' according to the literature [8]. For the substructures generated with top-down methods, the green colour implies a correct substructures and the red colour an incorrect substructure. The black colour represents bonds and substructures that are not covered. The black bond right next to a green or red substructures is the bond that is broken during fragmentation.

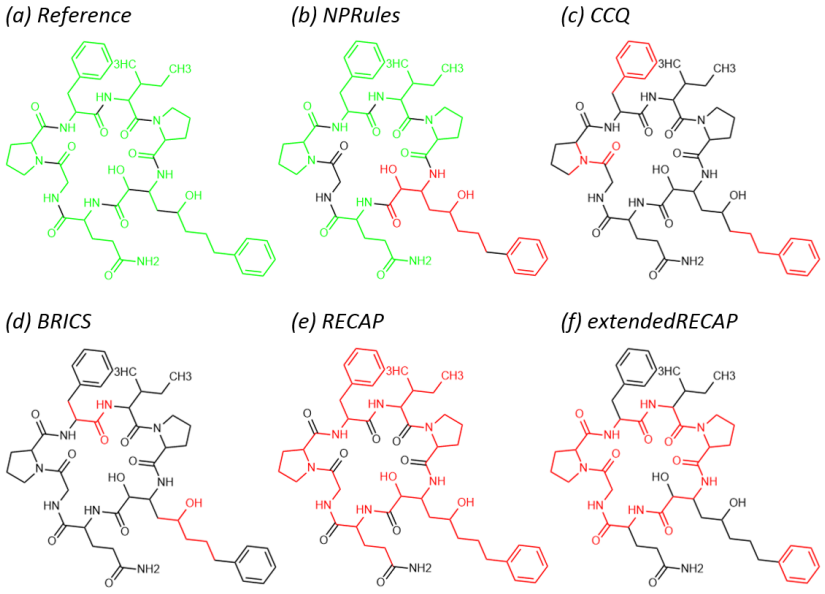


Figure 7. Nostophycin fragmented with NPRules, CCQ, BRICS, RECAP and extendedRECAP (b-f). The reference (a) shows the BGC precursors used as reference substructures [8]. The correct substructures are indicated by green and the incorrect substructures are indicated by red.

The reference structure (figure 7a) contains eight (one substructure is present twice) correct substructures. NPRules resulted in four correct substructures and all other methods did not result a single correct structure. Note that CCQ and BRICS both result substructures that can match the structure on several locations, however, this can be ignored due to the incorrectness of the substructures. The overview, with all substructure results generated with the 12 reference structures using the different fragmentation rule sets, can be found in SI Appendix, table S-27. The precision (correct generated substructures/ all generated substructures) and recall (correct generated substructures/ correct reference substructures) for all results using all top-down methods and reference structures are shown in figure 8, the standard deviations are represented by the error bars.

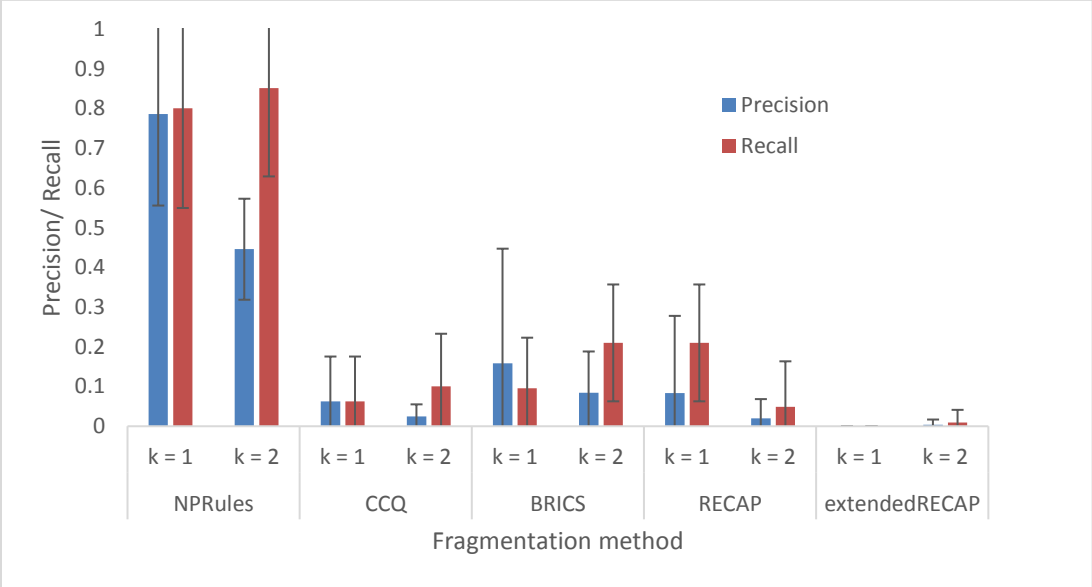


Figure 8. Precision and recall found for the BGC based validation results. The five different fragmentation rulesets were used to fragment 12 validation structures.

The NPRules results score the highest precision (0.79 for k=1 and 0.45 for k=2) and recall (0.80 for k=1 and 0.85 for k=2). BRICS and RECAP both result in a low precision and recall but still perform better than extendedRECAP. ExtendedRECAP does increase the substructure quantities compared to RECAP but, according this validation, decreases the quality. All fragmentation rulesets showed their own unique characteristic behaviours in substructure generation. CCQ resulted acyclic substructures and ignored crucial bonds while BRICS created many illogical (acyclic) substructures and caused some side chain cleavage. RECAP missed crucial bonds and excluded some side chains, extendedRECAP did not miss that many important bonds, but did exclude many side chains. The latter also explains why extendedRECAP scored the lowest precision and recall.

MS based validation results

The 12 reference mass spectra and the corresponding NP structures are provided in SI Appendix, section S1.4. The example structure shown in figure 9 is novobiocin. The green substructures in figure 9a are regarded as 'correct' according to the literature [23]. For the substructures generated with the different fragmentation rules, again, green implies a correct substructures and red implies an incorrect substructure. For each reference substructures the m/z value is shown and for the top-down generated substructures the exact mass is given (hydrogen atoms included).

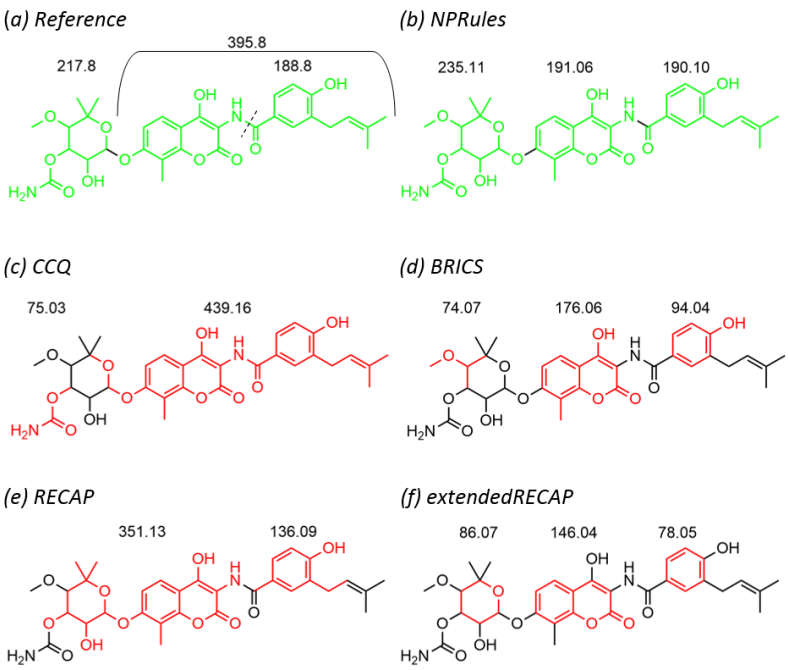


Figure 9. Novobiocin fragmented with NPRules, CCQ, BRICS, RECAP and extendedRECAP (b-f). The reference (a) shows the fragments, detected during MS analysis, used as reference substructures [23]. The correct substructures are indicated by green and the incorrect substructures are indicated by red. The values showed next to the substructures present in the reference structure represents the m/z value and for the others (b-f) the exact mass.

The MS reference spectra contains three major peaks (the heaviest ion excluded), see SI Appendix, section S1.4, Fig S-26. The fragment with an m/z of 395.8 has in itself another fragment, indicated by the green fragment on the right of the dotted line, with a m/z of 188.8 (figure 9a). The residual part, on the left of the dotted line, is considered a neutral loss of 207.0 (395.8 – 188.8). The substructures created with NPRules did have one 'exact' match with a difference of +1.30. This difference is legit as it is caused by the hydrogen loss due to ionization during MS analysis. The other match found was 'non-exact' with a difference of +17.31. Here, again one hydrogen atom was lost through MS analysis but also an oxygen atom (molecular mass 15.999) was missed by fragmentation with NPRules. The substructure we mention here is shown in figure 9b with a molecular mass of 235.11 g/mol. The last match was also 'non-exact', it missed an oxygen atom, and has a mass of 191.06 g/mol (figure 9b). Results of the other top-down methods for this example did not have any matches. For the MS validation result the number of exact and non-exact matches is determined for all 12 structures fragmented with the five different top-down fragmentation rules, the table with all results can be found in SI Appendix, table S-28. The results per ruleset are shown in table 2. The precision is calculated using both exact and non-exact matches as true positives.

Table 2. MS validation results using five different fragmentation rulesets.

Fragmentation ruleset	k	Total number of substructures	Exact match	Non-exact match	Precision
NPRules	1	33	17	3	0.606
	2	55	25	3	0.509
CCQ	1	22	1	2	0.136
	2	81	5	3	0.099
BRICS	1	29	3	2	0.172
	2	97	11	7	0.186
RECAP	1	23	2	3	0.217
	2	84	6	5	0.119
extendedRECAP	1	28	2	0	0.071
	2	112	6	7	0.116

The MS validation approach is more demanding than the BGC validation approach since more conditions should be met (correlation between m/z and exact mass and visual substructure recognition). Therefore, it does make sense that a decrease in precision was found using the MS validation method. NPRules creates the most exact matches, the number of non-exact matches is low/medium, and the precision again is the highest (0.606 for k=1 and 0.509 for k=2). BRICS and extendedRECAP (both k=2) do generate a sufficient number of substructure matches, 16 and 13, respectively. The high numbers of undesired substructures counteract the fragmentation method effects and causes a low precision.

Qualitative bottom-up results

The substructure's quality of the bottom-up results are ignored on purpose during validation since further method optimization is required (better pre-substructure generation and filter extension by implementing extra steps). Both pre-substructures generation methods (permutation and environmental) have their disadvantaged as mentioned in the methods and implementation section but do work and also the filter steps correctly eliminate substructures. To show the outcome of several parameter settings an example structure is fragmented (canonical SMILES: 'OC1C2CCC12') with two different sets of selected parameter settings. First a total of 18 pre-substructures was generated with the permutation method, tables 3 and 4 list the exact parameter settings values used during filtering. The final substructures are represented with a blue colour (figure 10 and 11).

Table 3. Parameter settings set 1.

Filter step	Parameter value	Number of substructures left
Molecular weight	≥ 40	16
Hetero atoms	≥ 0	16
Number of cycles	≥ 1	9
Abundancy	≥ 0 and ≤100	9
Number of bonds	≤ 2	5

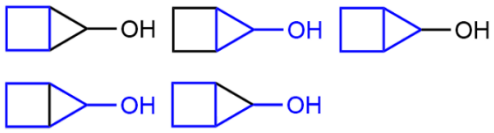


Figure 10. Final substructures left after filtering with parameter settings set 1. The substructures are represented by the blue colour.

Table 4. Parameter settings set 2.

Filter step	Parameter value	Number of substructures left
Molecular weight	≥ 0	18
Hetero atoms	≥ 1	8
Number of cycles	≥ 0	8
Abundancy	≥ 0 and ≤100	8
Number of bonds	≤ 3	7

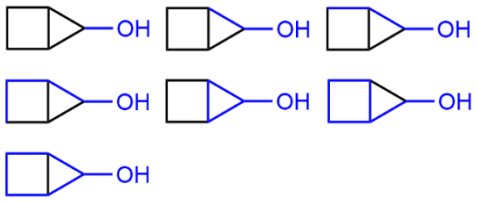


Figure 11. Final substructures left after filtering with parameter settings set 2. The substructures are represented by the blue colour.

The bottom-up method works according the given filter parameter values as can be observed in both figures. For example, figure 11 only shows substructures that include an oxygen atom, this is due to the parameter value of ≥ 1 for the hetero atom filter step. The bottom-up method, suitable for a wide range of different NP structures, can be used as starting point and inspiration for alternative structure fragmentation if the top-down method fails.

4. Discussion and Conclusion

The open-source NPDatabase was built for the storage of more than 320,000 unique NPs collected from other databases. It was important to cover the largest number of NP structures possible, in order to map the substructures found in BGC and mass spectra onto the correct structure. NPDatabase is user-friendly as data can be extracted with SQL. NP structures can be found with their canonical SMILES (generated with RDkit). Next to this straight forward approach, a selection of structures can be extracted for example based on molecular weight. The classification of the structures provides a selection of NP structures that share structural characteristics. NPDatabase can also be extended as new NP structures can be added.

First, the substructure quantity results were analyzed in order to examine the performance of the top-down and bottom-up methods. It is important to mention here that substructure quantities do not provide that much information. A fragmentation method can result a lot of substructures, but if those substructures are rubbish then the fragmentation method is useless. However, substructure quantities do give an indication whether a fragmentation method is able to fragment a NP structure, and in our case, NP structures of certain classes.

The top-down method using NPRules was validated with BGC and MS data, the other fragmentation rulesets (CCQ, BRICS, RECAP and extendedRECAP) were also taken into account. Now the substructure's quality was examined. Looking at BGC validation, NPRules results in the highest precision (0.79 for k=1 and 0.45 for k=2). and recall (0.80 for k=1 and 0.85 for k=2). NPRules also performed the best according to MS based validation with a precision of 0.606 for k=1 and a precision of 0.509 for k=2. Several challenges were faced during validation with MS data. The spectra were often not annotated and identified, variation between MS analyses methods (i.e. equipment and solvents) that result in different mass spectra for the same structure and the difference between noise and actual peaks is not always clear (which is also the reason for recall exclusion). The most important conclusion of the validation process is that NPRules does result more substructures similar to those in BGC and mass spectra than the other fragmentation rulesets do.

The bottom-up method is not optimal yet, but does offer new perspective with an opposite approach to generate substructures. This method is developed from scratch and the resulting substructures met the conditions set by the parameter values for the filter steps given by the user. The bottom-up results, using the method's current form, contain a lot of overlapping substructures that do not match those from BGC and MS data.

For now, the recommendation is to use the top-down method. NPRules is recommended for saccharide, ester and peptide classes but also for other NP structures that contain simple (implying connection through one bond) aromatic and cyclic compounds. For polypeptide/ amino acid structures, the advice is to use k=1 as that results single amino acids. It is also recommended to use a value of 2 for the parameter n (minimal number of atoms in a substructure). The BGC validation example for NPRules (figure 7b) would then have resulted in one extra correct substructure. According to the BGC validation results, looking at the NPRules results, k=1 causes a higher precision (0.343) and k=2 causes a slightly higher recall (+0.051). For the MS results, the precision is higher with 0.097 for k=1. Generally, k=1 is recommended for NPRules usage. If NPRules performs poor (low quantity results), other rulesets can be used instead. In that case BRICS or RECAP (RECAP not for polysaccharides) is recommended as it results high substructure quantities and better precision and recall than CCQ and extendedRECAP. For RECAP a value of 1 is recommended for parameter k. The differences between the k=1 and k=2 results for BRICS are lower, therefore the value choice of this parameter setting shall be left to the user.

The final goal, that goes beyond this project, is to generate all relevant substructures for every single NP structure stored in NPDatabase. These relevant substructures, linked to the structure they are fragmented from, then can be stored in the Substructure table. Ultimately, precursors from BGCs and mass fragments from mass spectra can be linked to the corresponding substructures, which provide further linkage to the structure they originated from. This structure then can be assigned as candidate structure to the BGC or the mass spectrum. An automated process that includes everything would be the most ideal. The recognition and annotation of NP substructures as described here promotes interpreting BGCs and MS data, which tackles a big problem in the genomics and metabolomics fields.

5. Future perspective

To achieve the final goal mentioned in the section above, first other steps have to be taken. To generate relevant substructures for all NPs in NPDatabase classification should be taken into account. NPRules can be used as base and extra fragmentation rules can be added which are more class specific. The NP structures now present in NPDatabase are subdivided in 499 classes and in 29 superclasses. Because of the latter, defining extra rules for each superclass is very reasonable. Note that knowledge about SMARTS patterns is required in order to define new fragmentation rules.
The bottom-up method also has to be optimized, a method to generate all possible pre-substructures with a short execution time should be found. Also extra filters should be added, for example a filter to exclude overlapping substructures and a filter that excludes substructures with missing side chains. If the bottom-up method is optimized, it can be used to generate substructures for more challenging (super)classes, like lactones having complex cyclic structures which are difficult to fragment with the top-down method.

6. Footnotes

NPDatabase can be downloaded at:
https://www.dropbox.com/s/qumnikhiaszrwjh/ NPDatabase.sqlite?dl=0

The instructions to install molBLOCKS, the NPRules text file and the python scripts for the bottom-up method (permutation and environmental) can be found at:
https://github.com/SamStokman/NPThesis

7. References

1. Katz, L. and R.H. Baltz, *Natural product discovery: past, present, and future.* Journal of industrial microbiology & biotechnology, 2016. **43**(2-3): p. 155-176.
2. Newman, D.J. and G.M. Cragg, *Natural products as sources of new drugs from 1981 to 2014.* Journal of natural products, 2016. **79**(3): p. 629-661.
3. Cimermancic, P., et al., *Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters.* Cell, 2014. **158**(2): p. 412-421.
4. Amos, G.C., et al., *Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality.* Proceedings of the National Academy of Sciences, 2017: p. 201714381.
5. Peisl, L., E.L. Schymanski, and P. Wilmes, *Dark matter in host-microbiome metabolomics: tackling the unknowns–a review.* Analytica chimica acta, 2018. **1037**: p. 13-27.
6. Wang, M., et al., *Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking.* Nature biotechnology, 2016. **34**(8): p. 828.
7. van der Hooft, J.J., et al., *Topic modeling for untargeted substructure exploration in metabolomics.* Proc Natl Acad Sci U S A, 2016. **113**(48): p. 13738-13743.
8. Fewer, D.P., et al., *Nostophycin biosynthesis is directed by a hybrid PKS-NRPS in the toxic cyanobacterium Nostoc sp. 152.* Applied and environmental microbiology, 2011: p. AEM. 05993-11.
9. Cacho, R.A., Y. Tang, and Y.-H. Chooi, *Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi.* Frontiers in microbiology, 2015. **5**: p. 774.
10. Greule, A., et al., *From a natural product to its biosynthetic gene cluster: a demonstration using polyketomycin from Streptomyces diastatochromogenes Tü6028.* Journal of visualized experiments: JoVE, 2017(119).
11. Khijwania, P.S., *Biosynthesis of a Few Members of Natural Enediynes (Biosynthesis of 10-Membered Enediynes , Calicheamicin γ1I)-Part-VI*, in *Bio-Organic Chemistry of Natural Enediyne Anticancer Antibiotivs (Web)*, I. Guwahati, Editor. 2013, NPTEL, A Project funded bu MHRD, Govt. of India.
12. Kim, C.-G., et al., *Biosynthesis of rubradirin as an ansamycin antibiotic from Streptomyces achromogenes var. rubradiris NRRL3061.* Archives of microbiology, 2008. **189**(5): p. 463-473.
13. Lai, J.R., et al., *A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection.* Proceedings of the National Academy of Sciences, 2006. **103**(14): p. 5314-5319.
14. Li, Y., et al., *Coordinative modulation of chlorothricin biosynthesis by binding of the glycosylated intermediates and end product to a responsive regulator ChlF1.* Journal of Biological Chemistry, 2016: p. jbc. M115. 695874.

15. Liu, X., et al., *Proteomic response of methicillin-resistant S. aureus to a synergistic antibacterial drug combination: a novel erythromycin derivative and oxacillin.* Scientific reports, 2016. **6**: p. 19841.
16. Luzhetskyy, A., A. Vente, and A. Bechthold, *Glycosyltransferases involved in the biosynthesis of biologically active natural products that contain oligosaccharides.* Molecular BioSystems, 2005. **1**(2): p. 117-126.
17. Oppegard, L.M., et al., *In vivo and in vitro patterns of the activity of simocyclinone D8, an angucyclinone antibiotic from Streptomyces antibioticus.* Antimicrobial agents and chemotherapy, 2009. **53**(5): p. 2110-2119.
18. Schmartz, P.C., et al., *Bis-chlorination of a hexapeptide–PCP conjugate by the halogenase involved in vancomycin biosynthesis.* Organic & biomolecular chemistry, 2014. **12**(30): p. 5574-5577.
19. Baars, O. and D.H. Perlman, *Small Molecule LC-MS/MS Fragmentation Data Analysis and Application to Siderophore Identification*, in *Applications from Engineering with MATLAB Concepts*. 2016, InTech.
20. Damsten, M., et al., *LC-tandem MS detection of covalent binding of acetaminophen to human serum albumin.* Drug metabolism and disposition, 2007.
21. De Vijlder, T., et al., *A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation.* Mass spectrometry reviews, 2018. **37**(5): p. 607-629.
22. Ding, W., et al., *Biosynthetic investigation of phomopsins reveals a widespread pathway for ribosomal natural products in Ascomycetes.* Proceedings of the National Academy of Sciences, 2016. **113**(13): p. 3521-3526.
23. Inoue, K., et al., *LC-MS/MS and centrifugal ultrafiltration method for the determination of novobiocin in chicken, fish tissues, milk and human serum.* Journal of Chromatography B, 2009. **877**(4): p. 461-464.
24. Luzzatto-Knaan, T., et al., *Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae.* eLife, 2017. **6**: p. e24214.
25. Posocco, B., et al., *A new high-performance liquid chromatography-tandem mass spectrometry method for the determination of paclitaxel and 6α-hydroxy-paclitaxel in human plasma: Development, validation and application in a clinical pharmacokinetic study.* PloS one, 2018. **13**(2): p. e0193500.
26. Sandermann, H., et al., *A New Intermediate in the Mineralization of 3, 4-Dichloroaniline by the White Rot FungusPhanerochaete chrysosporium.* Applied and environmental microbiology, 1998. **64**(9): p. 3305-3312.
27. Tracqui, A. and B. Ludes, *HPLC-MS for the determination of sildenafil citrate (Viagra®) in biological fluids. Application to the salivary excretion of sildenafil after oral intake.* Journal of Analytical toxicology, 2003. **27**(2): p. 88-94.
28. Trösken, E.R., et al., *Quantitation of lanosterol and its major metabolite FF-MAS in an inhibition assay of CYP51 by azoles with atmospheric pressure photoionization based LC-MS/MS.* Journal of the American Society for Mass Spectrometry, 2004. **15**(8): p. 1216-1221.
29. Zhu, L., et al., *Simultaneous determination of methylephedrine and noscapine in human plasma by liquid chromatography–tandem mass spectrometry.* Journal of Chromatography B, 2005. **820**(2): p. 175-182.
30. Degen, J., et al., *On the Art of Compiling and Using'Drug-Like'Chemical Fragment Spaces.* ChemMedChem: Chemistry Enabling Drug Discovery, 2008. **3**(10): p. 1503-1507.
31. Heikamp, K., et al., *Exhaustive sampling of the fragment space associated to a molecule leading to the generation of conserved fragments.* Chemical biology & drug design, 2018. **91**(3): p. 655-667.
32. Ghersi, D. and M. Singh, *molBLOCKS: decomposing small molecule sets and uncovering enriched fragments.* Bioinformatics, 2014. **30**(14): p. 2081-2083.
33. Liu, T., et al., *Break down in order to build up: decomposing small molecules for fragment-based drug design with e MolFrag.* Journal of chemical information and modeling, 2017. **57**(4): p. 627-631.

Creating a Natural Product Database and Generating Molecular Substructures

Supporting Appendix

Contents

Section S1. Fragmentation methods

This section contains supplementary information about the development and the validation of the top-down and bottom-up fragmentation methods.

Section S1.1 Top-down fragmentation with NPRules

In total 45 rules, listed in table S-1, were defined which together represent the new fragmentation ruleset NPRules. NPRules is highly suitable for NPs from saccharide, ester and peptide classes.

Table S-1. NPRules defined with SMARTS patterns.

SMARTS pattern	**Rule description**
[c]!@[c]	aro_c-aro_c
[c]!@[n]	aro_c-aro_n
[c]!@[C$(C[!Ch3,Cl,S])]	aro_c-ali_C_Cl_S
[c]!@[C$(C(@C)@C)]	aro_c-cyclic_ali_c
[c]!@[C$(C(@c)@C)]	aro_c-cyclic_ali&aro_c
[c]!@[C$(C!@C!@C!@C!@C)]	aro_c-pentane_chain
[c]!@[C$(C!@=C!@-C!@=C!@-C)]	aro_c-pentene_chain
[c]!@[C$(C!@N@C)]	aro_c-specific1
[c]!@[C$(C=C@C)]	aro_c- specific2
[c]!@[S$(S[c,C])]	aro_c-S_ali&aro_c
[C$(C(@C)@C)]!@[n]	ali_cyclic_C-aro_n
[C$(C(@C)@C)]!@[C$(C(@C)@C)]	ali_cyclic_C-ali_cyclic_C
[C$(C(@C)@C)]!@[C$(C(@C)@O)]	ali_cyclic_C-ali_cyclic_C&O
[C$(C(@C)@C)]!@[C$(C(!@C@O)!@C)]	ali_cyclic_C-specific1
[C$(C(@C)@C)]!@[C$(C(!@C!@C)!@C)]	ali_cyclic_C-specific2
[C$(C(@C)@C)]!@[C$(C!@C!@C!@C!@C)]	ali_cyclic_C-pentane_chain
[C$(C(@C)@C)]!@[C$(C!@=C!@-C!@=C!@-C)]	ali_cyclic_C-pentene_chain
[C$(C(@[C,c])@O)]!@[C$(C(@[C,c])@N)]	cyclic_ali_C&O-cyclic_ali_C&N
[C$(C(@O)=@C@C)]!@[C$(C=!@C!@C)]	specific3-specific4
[C$(C(c)=O)]-[O$(OC@C)]	specific5-specific6
[c]-[$(*1O****1)]	aro_c-sugar_wildcards1
[c]-[$(*1**O**1)]	aro_c-sugar_wildcards3
[c]-[O$(OC1OCCCC1)]	aro_c-O_sugar1
[c]-[O$(OCC1OCCCC1)]	aro_c-O_sugar1.1
[O$(OC1OCCCC1)]-[C$(C1COCCC1)]	O_sugar1-sugar2
[O$(OC1OCCCC1)]-[C$(C1CCOCC1)]	O_sugar1-sugar3
[O$(OC1OCCCC1)]-[C$(C!@C@C)]	O_sugar1-acylcic_C_cyclic_C
[O$(OC1OCCCC1)]-[C$(C(@C)@C)]	O_sugar1-cyclic_ali_c
[O$(OC1OCCC1)]-[C$(C1OCCC1)]	O_fructose1-fructose
[NH1,NH0]-[CH0$(C=!@O)]	amide=O
[NH0]=[CH0$(C-O)]	amide-OH
[C$(C(=O)c)]!@[O$(O!@C!@C@C)]	ester-aro&specific
[C$(C(=O)c)]!@[O$(Oc)]	ester-aro&aro
[C$(C(=O)c)]!@[O$(OC(@C)(@C))]	ester-aro&ali
[C$(C(=O)c)]!@[O$(O!@C!@C@C)]	ester-aro&specific
[C$(C(=O)C(@C)@C)]!@[O$(Oc)]	ester-ali&aro
[C$(C(=O)C(@C)@C)]!@[O$(OC(@C)(@C))]	ester-ali&ali
[C$(C(=O)!@-C!@=C!@-C!@=C)]!@[O$(OC(@C)(@C))]	ester-pentene_chain&ali
[C$(C(=O)!@-C!@-C!@-C!@-C)]!@[O$(OC(@C)(@C))]	ester pentane_chain&ali
[C$(C(=O)!@-C!@=C!@-C!@=C)]!@[O$(Oc)]	ester-pentene_chain&aro
[C$(C(=O)!@-C!@-C!@-C!@-C)]!@[O$(Oc)]	ester-pentane_chain&aro
[C$(C(=O)C(@C)@C)]!@[O$(O!@-C!@-C!@-C!@-C)]	ester-ali&pentane_chain
[C$(C(=O)C(@C)@C)]!@[O$(O!@-C!@=C!@-C!@=C)]	ester-ali&pentene_chain
[C$(C(=O)c)]!@[O$(O!@-C!@-C!@-C!@-C)]	ester-aro&pentane_chain
[C$(C(=O)c)]!@[O$(O!@-C!@=C!@-C!@=C)]	ester-aro&pentene_chain

Section S1.2 Bottom-up fragmentation; pre-substructure generation

The execution time for pre-substructure generation with varying SMILES strings lengths (2 to 28) was determined for the permutation and the environmental approach. The results are shown in figure S-2.

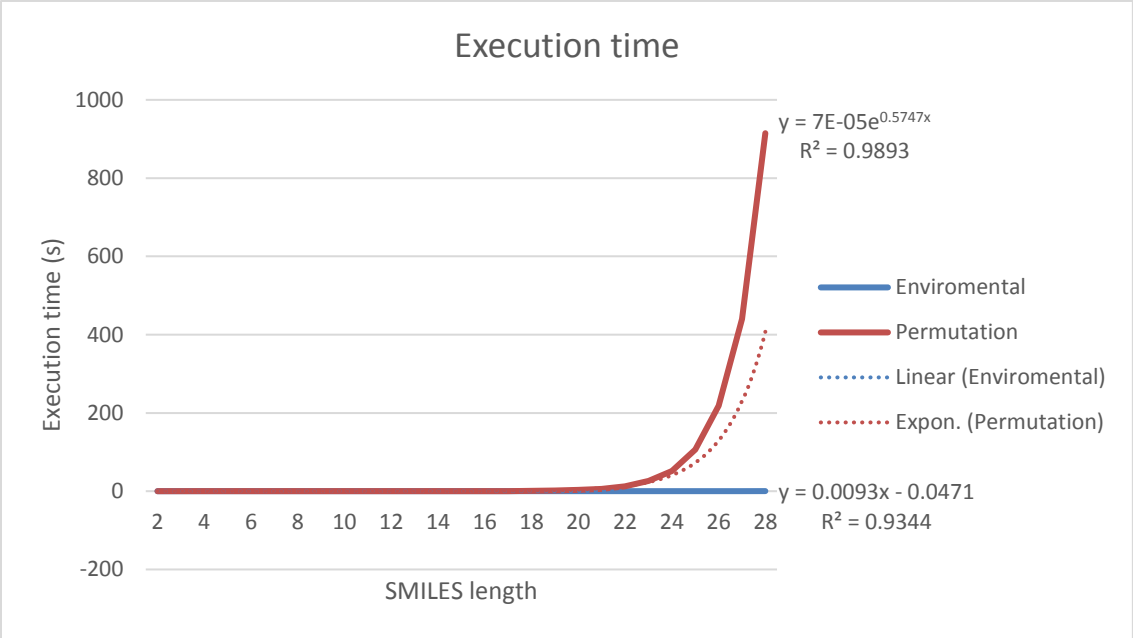


Figure S-2. Execution time permutation vs. environmental pre-substructure generation.

The relationship between the execution time and SMILES string length using the permutation approach is exponential with a R^2 of 0.9893. Therefore, only NP structures with a maximum SMILES string length of 18 are recommended. The environmental approach shows a linear relationship with a R^2 of 9.344 and is recommended for longer input SMILES.

Section S1.3 BGC validation structures

In total 12 reference structures (figure S-3 – S-14) originating from BGC data were used for method validation. The substructures showed here are considered as 'correct' substructures, and used for comparison to the substructures generated with all top-down methods.

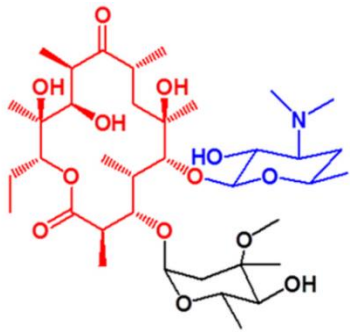


Figure S-3. Validation structure erythomycin[13].

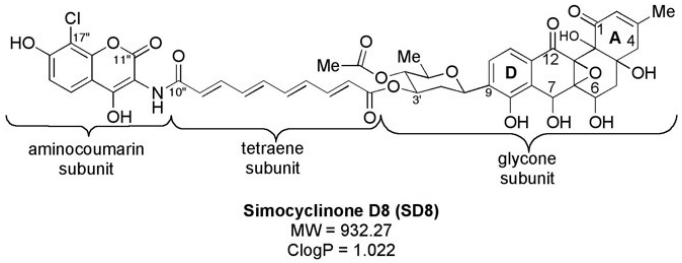


Figure S-4. Validation structure simocyclinone [16].

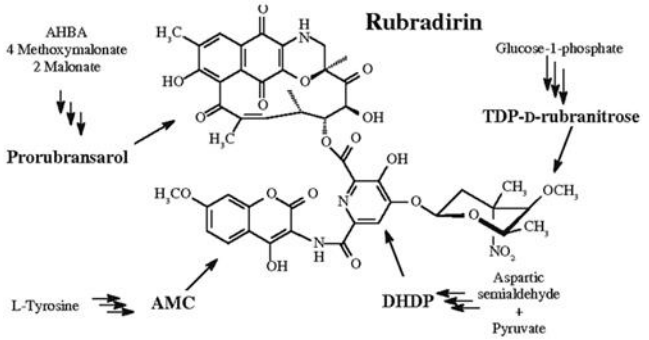


Figure S-5. Validation structure rubradirin [17].

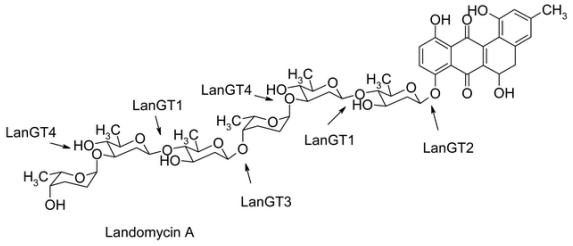


Figure S-6. Validation structure landomycin A[3]

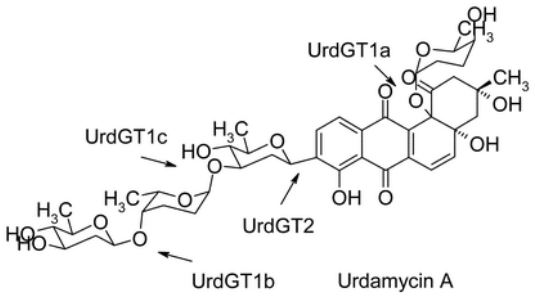


Figure S-7. Validation structure urdamycin A[3]

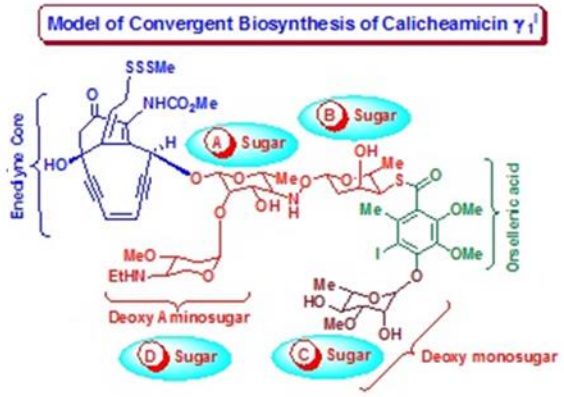


Figure S-8. Validation structure calicheamicin[20]

Figure S-9. Validation structure chlorothricin[10].

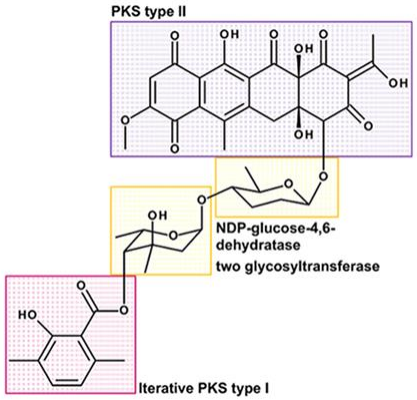


Figure S-10. Validation structure polyketomycin[4].

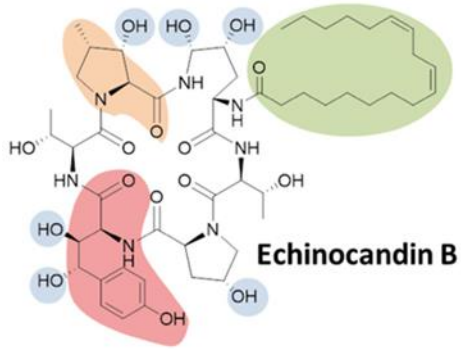


Figure S-11. Validation structure echinocandin B[2].

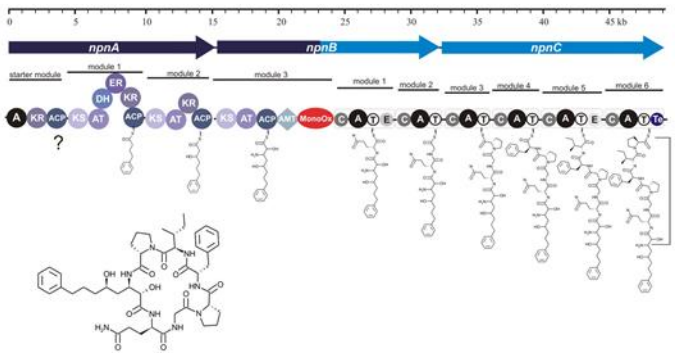


Figure S-12. Validation structure nostophycin[9].

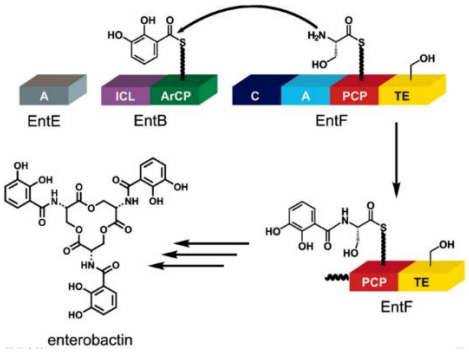


Figure S-13. Validation structure enterobactin[19].

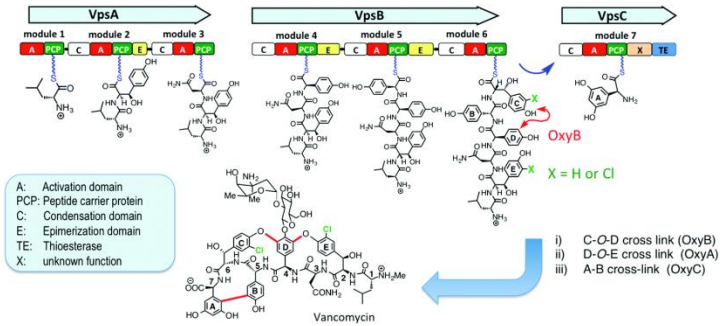


Figure S-14. Validation structure vancomycin[5].

Section S1.4 MS validation structures

In total 12 reference structures (figure S-15 – S-24) represented by MS spectra were used for method validation. The substructures showed here are considered as 'correct' substructures, and used for comparison to the substructures generated with all top-down methods.

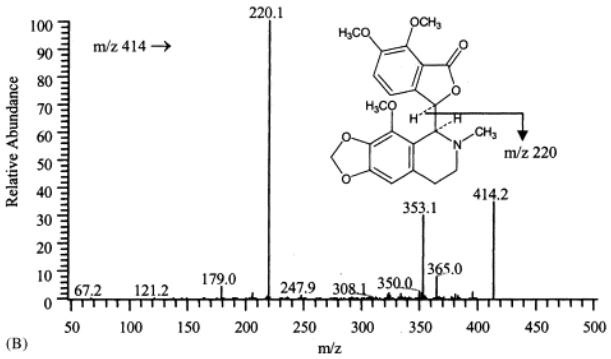


Figure S-15. Validation structure noscapine[7]

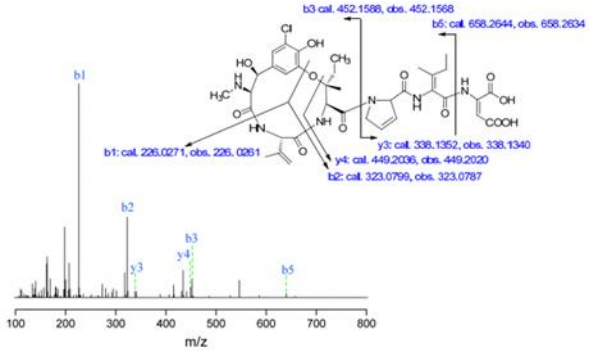


Figure S-16. Validation structure phomopsin A[13].

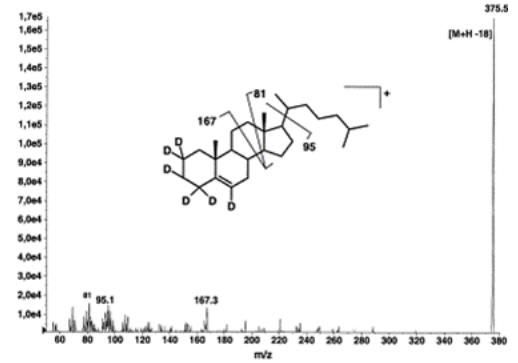


Figure S-17. Validation structure cholesterol d6[15].

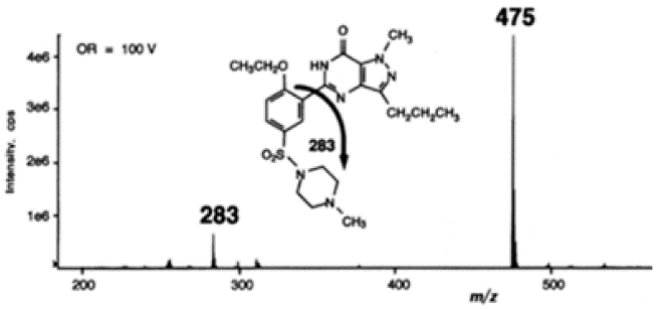


Figure S-182. Validation structure sildenafil[6].

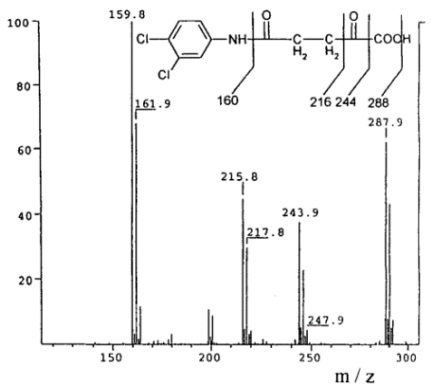


Figure S-19. Validation structure DCAX[12]

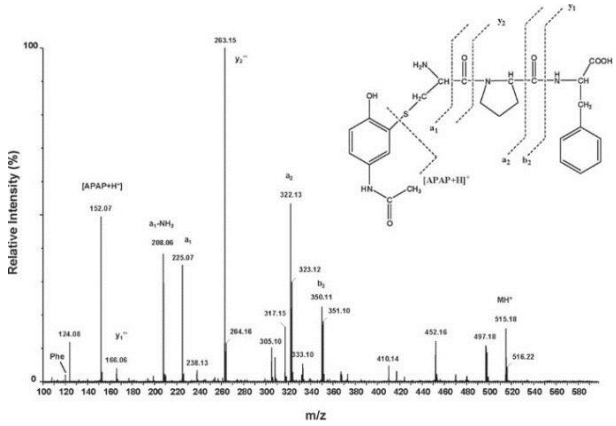


Figure S-201. Validation structure Reactive metabolite of Acetaminophen (NAPQI)[9]

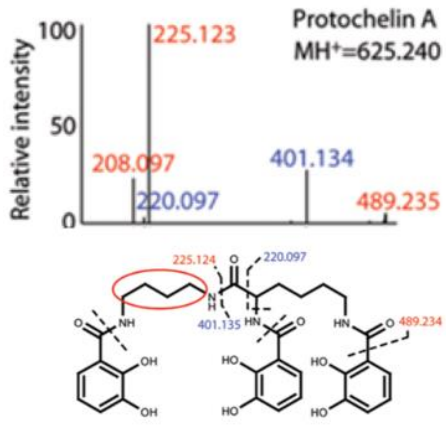


Figure S-21. Validation structure protochelin[1].

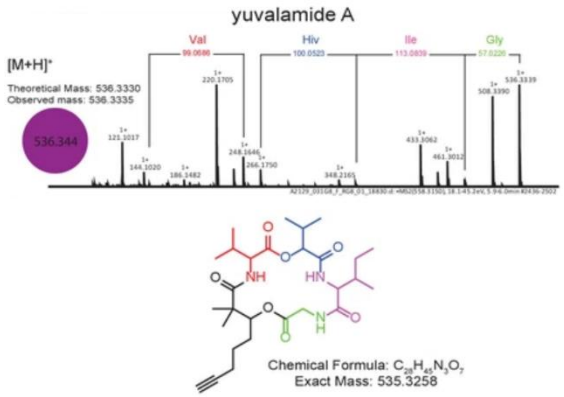


Figure S-22. Validation structure yuvalamide A[14]

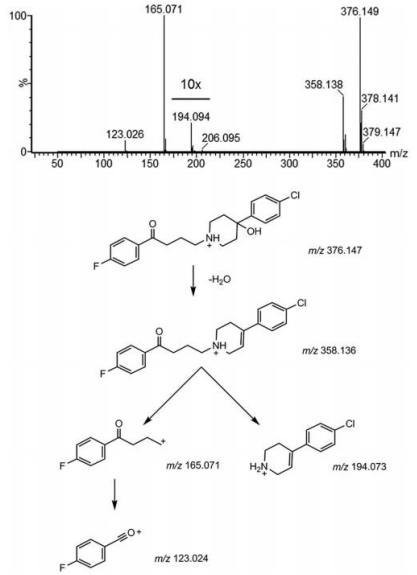


Figure S-23. Validation structure haloperidol[11].

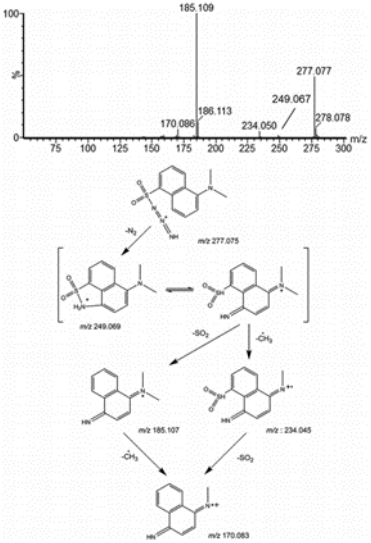


Figure S-24. Validation structure azide derivade[9].

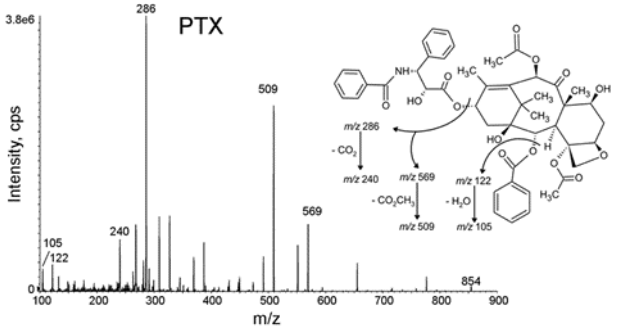


Figure S-25. Validation structure paclitaxel[18].

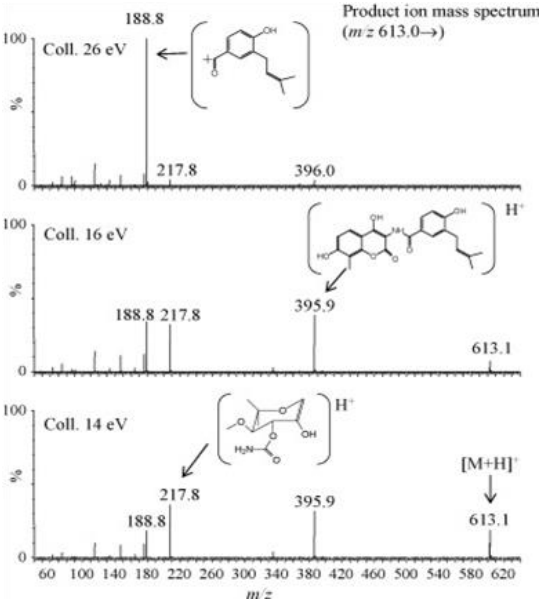


Figure S-263. Validation structure novobiocin[8].

Section S2. Supporting Results

This section provides the supplementary tables that contain the validation results based on BGC and MS data.

Section S2.1 BGC validation results

The substructure results generated with the 12 BGC reference structures using the different fragmentation rulesets are showed in table S-27. 'Total' is the total number of generated substructures and 'Correct' represents the number of substructures considered correct. Also the number of correct substructures in the reference structure is included.

Table S-27. Validation results BGC data

			NP structure											
			Erythromycin	Simocyclinone	Rubradirin	Landomycin	Urdamycin	Calicheamicin	Chlorothricin	Polyketomycin	Echinocandin B	Nostophycin	Enterobactin	Vancomycin
Number of correct reference substructures			3	3	4	4	5	6	4	4	3	8	2	9
Ruleset	**k**													
NPRules	1	Total	3	4	4	4	5	5	4	4	6	6	2	6
		Correct	3	2	4	4	5	3	4	4	3	4	1	4
	2	Total	5	7	7	9	9	10	7	7	13	14	3	14
		Correct	3	3	4	4	5	4	4	4	3	4	1	5
CCQ	1	Total	2	4	4	2	2	6	4	4	4	3	1	3
		Correct	0	0	1	0	0	0	1	1	0	0	0	0
	2	Total	15	17	18	13	15	29	18	17	17	15	3	18
		Correct	0	0	1	0	0	0	1	1	1	1	0	0
BRICS	1	Total	2	3	5	1	5	4	4	3	3	3	1	6
		Correct	0	1	1	1	1	0	0	0	0	0	0	1
	2	Total	14	12	20	11	13	27	20	13	12	8	3	23
		Correct	0	1	1	1	1	1	1	1	0	3	0	4
RECAP	1	Total	1	2	3	0	1	2	2	2	3	1	1	1
		Correct	0	1	0	0	0	0	0	1	0	0	0	0
	2	Total	5	7	13	0	4	2	9	10	13	10	2	3
		Correct	0	1	0	0	0	0	0	1	0	0	0	0
Extended-RECAP	1	Total	2	3	4	2	2	3	3	3	4	2	2	3
		Correct	0	0	0	0	0	0	0	0	0	0	0	0
	2	Total	20	18	22	11	12	26	16	19	17	9	5	22
		Correct	0	0	0	0	0	0	0	0	0	0	0	1

Section S2.3 MS validation results

The substructure results for MS validation, using 12 input structures, are showed in table S-28. The total number of substructures, the exact and the non-exact matches can be observed.

Table S-28. Validation results MS data

Ruleset	**k**		**NP structure**											
			Noscapine	Phomopsin A	Cholesterol d6	Sildenafil	DCAX	NAPQI	Protochelin B	Haloperidol	Azide derivate	Yuvalamide A	Paclitaxel	Novobiocin
NPRules	1	Total	2	5	2	3	2	4	3	2	2	3	2	3
		Exact match	2	2	2	1	2	3	2	0	0	2	0	1
		Non-exact match	0	0	0	0	0	0	0	0	0	0	1	2
	2	Total	2	10	2	5	2	8	7	2	2	6	4	5
		Exact match	2	5	2	2	2	6	3	0	0	2	0	1
		Non-exact match	0	0	0	0	0	0	0	1	0	0	2	3
CCQ	1	Total	2	4	1	1	1	3	1	2	0	1	4	2
		Exact match	1	0	0	0	0	0	0	0	0	0	0	0
		Non-exact match	0	0	0	0	0	0	0	1	0	0	1	0
	2	Total	3	17	1	4	3	10	5	5	0	9	17	7
		Exact match	2	0	0	0	1	1	1	0	0	0	0	0
		Non-exact match	0	0	0	0	0	1	0	1	0	0	1	0
BRICS	1	Total	1	3	2	2	2	4	2	3	1	3	3	3
		Exact match	0	0	1	0	1	0	1	0	0	0	0	0
		Non-exact match	0	1	0	0	1	0	0	0	0	0	0	0
	2	Total	6	14	3	6	4	10	7	7	1	12	14	13
		Exact match	0	2	2	1	2	1	2	1	0	0	0	0
		Non-exact match	0	1	0	0	2	1	0	0	0	1	1	1
RECAP	1	Total	1	2	0	3	2	3	3	2	1	1	3	2
		Exact match	0	0	0	0	1	0	1	0	0	0	0	0
		Non-exact match	0	0	0	0	0	1	1	1	0	0	0	0
	2	Total	6	13	0	8	6	11	9	4	3	6	11	7
		Exact match	0	0	0	1	1	1	2	1	0	0	0	0
		Non-exact match	0	0	0	0	0	1	1	1	0	0	1	0
Extended-RECAP	1	Total	2	3	2	3	2	4	3	2	1	1	2	3
		Exact match	0	0	0	0	1	0	1	0	0	0	0	0
		Non-exact match	0	0	0	0	0	0	0	0	0	0	0	0
	2	Total	6	16	7	9	5	13	9	9	3	7	13	15
		Exact match	1	0	0	0	1	2	2	0	0	0	0	0
		Non-exact match	0	1	0	0	1	1	0	2	1	0	1	0

References

1. Baars, O. and D.H. Perlman, *Small Molecule LC-MS/MS Fragmentation Data Analysis and Application to Siderophore Identification*, in *Applications from Engineering with MATLAB Concepts*. 2016, InTech.
2. Cacho, R.A., Y. Tang, and Y.-H. Chooi, *Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi.* Frontiers in microbiology, 2015. **5**: p. 774.
3. Luzhetskyy, A., A. Vente, and A. Bechthold, *Glycosyltransferases involved in the biosynthesis of biologically active natural products that contain oligosaccharides.* Molecular BioSystems, 2005. **1**(2): p. 117-126.
4. Greule, A., et al., *From a natural product to its biosynthetic gene cluster: a demonstration using polyketomycin from Streptomyces diastatochromogenes Tü6028.* Journal of visualized experiments: JoVE, 2017(119).
5. Schmartz, P.C., et al., *Bis-chlorination of a hexapeptide–PCP conjugate by the halogenase involved in vancomycin biosynthesis.* Organic & biomolecular chemistry, 2014. **12**(30): p. 5574-5577.
6. Tracqui, A. and B. Ludes, *HPLC-MS for the determination of sildenafil citrate (Viagra®) in biological fluids. Application to the salivary excretion of sildenafil after oral intake.* Journal of Analytical toxicology, 2003. **27**(2): p. 88-94.
7. Zhu, L., et al., *Simultaneous determination of methylephedrine and noscapine in human plasma by liquid chromatography–tandem mass spectrometry.* Journal of Chromatography B, 2005. **820**(2): p. 175-182.
8. Inoue, K., et al., *LC-MS/MS and centrifugal ultrafiltration method for the determination of novobiocin in chicken, fish tissues, milk and human serum.* Journal of Chromatography B, 2009. **877**(4): p. 461-464.
9. Damsten, M., et al., *LC-tandem MS detection of covalent binding of acetaminophen to human serum albumin.* Drug metabolism and disposition, 2007.
10. Li, Y., et al., *Coordinative modulation of chlorothricin biosynthesis by binding of the glycosylated intermediates and end product to a responsive regulator ChlF1.* Journal of Biological Chemistry, 2016: p. jbc. M115. 695874.
11. De Vijlder, T., et al., *A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation.* Mass spectrometry reviews, 2018. **37**(5): p. 607-629.
12. Sandermann, H., et al., *A New Intermediate in the Mineralization of 3, 4-Dichloroaniline by the White Rot FungusPhanerochaete chrysosporium.* Applied and environmental microbiology, 1998. **64**(9): p. 3305-3312.
13. Liu, X., et al., *Proteomic response of methicillin-resistant S. aureus to a synergistic antibacterial drug combination: a novel erythromycin derivative and oxacillin.* Scientific reports, 2016. **6**: p. 19841.
14. Luzzatto-Knaan, T., et al., *Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae.* eLife, 2017. **6**: p. e24214.
15. Trösken, E.R., et al., *Quantitation of lanosterol and its major metabolite FF-MAS in an inhibition assay of CYP51 by azoles with atmospheric pressure photoionization based LC-MS/MS.* Journal of the American Society for Mass Spectrometry, 2004. **15**(8): p. 1216-1221.
16. Oppegard, L.M., et al., *In vivo and in vitro patterns of the activity of simocyclinone D8, an angucyclinone antibiotic from Streptomyces antibioticus.* Antimicrobial agents and chemotherapy, 2009. **53**(5): p. 2110-2119.
17. Kim, C.-G., et al., *Biosynthesis of rubradirin as an ansamycin antibiotic from Streptomyces achromogenes var. rubradiris NRRL3061.* Archives of microbiology, 2008. **189**(5): p. 463-473.
18. Posocco, B., et al., *A new high-performance liquid chromatography-tandem mass spectrometry method for the determination of paclitaxel and 6α-hydroxy-paclitaxel in human plasma: Development, validation and application in a clinical pharmacokinetic study.* PloS one, 2018. **13**(2): p. e0193500.
19. Lai, J.R., et al., *A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection.* Proceedings of the National Academy of Sciences, 2006. **103**(14): p. 5314-5319.
20. Khijwania, P.S., *Biosynthesis of a Few Members of Natural Enediynes (Biosynthesis of 10-Membered Enediynes , Calicheamicin γ1I)-Part-VI*, in *Bio-Organic Chemistry of Natural Enediyne Anticancer Antibiotivs (Web)*, I. Guwahati, Editor. 2013, NPTEL, A Project funded bu MHRD, Govt. of India.