A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products


ABSTRACT: This report describes the first application of the novel NMR-based machine learning tool “Small Molecule Accurate Recognition Technology” (SMART 2.0) for mixture analysis and subsequent accelerated discovery and characterization of new natural products. The concept was applied to the extract of a filamentous marine cyanobacterium known to be a prolific producer of cytotoxic natural products. This environmental Symploca extract was roughly fractionated, and then prioritized and guided by cancer cell cytotoxicity, NMR-based SMART 2.0, and MS²-based molecular networking. This led to the isolation and rapid identification of a new chimeric swinholide-like macrolide, symplocolide A, as well as the annotation of swinholide A, samholides A−I, and several new derivatives. The planar structure of symplocolide A was confirmed to be a structural hybrid between swinholide A and luminoide B by 1D/2D NMR and LC-MS² analysis. A second example applies SMART 2.0 to the characterization of structurally novel cyclic peptides, and compares this approach to the recently appearing “atomic sort” method. This study exemplifies the revolutionary potential of combined traditional and deep learning-assisted analytical approaches to overcome longstanding challenges in natural products drug discovery.

Natural products (NPs) of terrestrial and marine organisms have been a highly valuable source of leads for biomedical applications. About 50% of FDA approved drugs can trace their origin to NPs, and notably, marine NPs have proportionately produced a much higher success rate than other sources of drug leads (e.g., 19 approved marine derived or inspired agents are on the market today). Filamentous marine cyanobacteria are especially rich in structurally diverse and biologically active NPs, such as those showing differential cytotoxicity against various cancer cell lines. In this regard, the genus Symploca is a prolific producer of cytotoxic NPs such as dolastatin 10, symplomamide A, and others. 

Drug discovery and development is expensive ($2 billion on average), time-consuming (13−15 years from concept to market), and risky (clinical success rate of 12%). Informatic tools are therefore being developed to make the discovery of new leads more efficient, to repurpose known agents, to target new metabolites on the basis of genomic analysis, to rapidly reveal mechanisms of action, and to optimize the pharmaceutical properties of drug leads. For example, the emerging research area of genome mining featuring bioinformatic tools such as AntiSMASH 5.0 and BiGSCAPE provide an orthogonal approach to NP discovery, although an approved drug discovered by this approach has not yet been marketed. The rapid identification of molecular structure, either known or new, is a significant challenge in NP discovery, and a variety of partial solutions are emerging using disparate analytical approaches.

For example, liquid chromatography−tandem mass spectrometry-based (LC-MS/MS) dereplication tools, such as the Global Natural Products Social molecular networking (GNPS; http://gnps.ucsd.edu), MS2LDA, and SIRIUS² represent paradigm shifts in NP research. These new tools are facilitating the targeted isolation of new NPs as well as rapidly dereplicating known ones. Nevertheless, unambiguous identification of new NPs still requires isolation and characterization by NMR spectroscopy to comply with the minimum standards for novel metabolite annotation. Recent efforts have focused on combining various NMR techniques with mass spectrometry and in silico databases. 

Previously, we reported a novel approach that involved training a deep convolutional neural network (CNN) of Siamese architecture with 2048 1H−13C HSQC spectra mined from the Supporting Information sections of the Journal of Natural Products. This trained system was used to analyze new spectra, place them within 10-dimensional SMART cluster space, and accelerate the structure elucidation of a series of cyclic lipopeptides.

Encouraged, we subsequently developed automated pipelines to produce constructed HSQC spectra from data tables as well as predicted HSQC spectra from published structures. SMART 2.0 was trained on 25 434 HSQC spectra from NPs of...
the JEOL database (https://www.j-resonance.com/en/nmrdb/). The spectra were mapped into a 180-dimensional embedding space using a CNN (SqueezeNet).\(^\text{36}\) We mapped an additional 27,642 spectra that were computed using the ACD Laboratories predictor to create the HSQC spectra of metabolites from these fractions using GNPS. The latter revealed a highly abundant and distinct cluster of features with the major m/z 1395.9 in active fraction H4 (Figure 1e). Isolation and structure elucidation of this compound led to the discovery of a new cytotoxic macrolide \(1\) (Scheme 1), named symplocolide A (1), representing a structural chimera between swinholide A (4) and luminaolide B (3).\(^\text{43}\) Analysis of the MS\(^2\) spectra of metabolites from these fractions using GNPS led to annotation of swinholide A (4), the glycosylated samholides A–I (5–13, Scheme 1 Figures S12–S22), as well as several putatively new analogues swinholides L and M, samholides J–M, and symplocolides B–D (Figures S23–S30).

Here, we demonstrate this unique cheminformatic tool to automatically characterize a complex NP from a cyanobacterial extract mixture for the first time. Resultantly, the extract was fractionated and the isolation of novel NPs undertaken guided by NMR-based SMART mixture analysis, MS\(^2\) molecular networking, and cytotoxicity against NCI-H460 human lung cancer cells in vitro. This led to the discovery of a new swinholide class of NP, named “symplocolide A” (compound 1, Scheme 1, Scheme S1), as well as the annotation of several known compounds in this structural class.

**Scheme 1. Structures of Symplocolide A (1), Luminaolide (2),\(^\text{32}\) Luminaolide B (3),\(^\text{33}\) Swinholide A (4), and Samholides A–I (5–13)\(^\text{44,44}\)**


```
<table>
<thead>
<tr>
<th>Structure</th>
<th>Chemical Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symplocolide A</td>
<td>R_1=CH_2; R_2=CH_3</td>
</tr>
<tr>
<td>Luminaolide B</td>
<td>R_1=CH_2; R_2=CH_3</td>
</tr>
<tr>
<td>Swinholide A</td>
<td>R_1=CH_2; R_2=CH_3</td>
</tr>
<tr>
<td>Samholide A</td>
<td>R_1=CH_2; R_2=CH_3</td>
</tr>
</tbody>
</table>
```

“All compounds except 2 and 3 were detected in this study.”

Tufts of a filamentous marine cyanobacterium, morphologically identified as *Symplocia* sp., were collected near American Samoa. The preserved collection was repetitively extracted (2/1 CH\(_2\)Cl\(_2\)/MeOH) and fractionated using vacuum liquid chromatography (VLC) to obtain nine fractions of increasing polarity (A–I). All fractions and the crude extract were screened for cytotoxicity to H-460 human lung cancer cells.\(^\text{38}\) Besides the crude extract, only fraction H showed strong cytotoxicity, while fraction I was moderately active (Figures 1a and S3). For initial SMART 2.0 analysis of the most cytotoxic crude fraction H, a NUS-ASAP-HSQC\(^\text{39–41}\) was recorded in 13 min on about 1 mg material in a 1.7 mm TCI MicroCryoProbe (599.10 MHz) to obtain the correlations of the major components of that fraction. After autopick picking of the 45 most intense features, 11 out of the top 12 structures annotated by SMART 2.0 were macrolides including the known cyanobacterial metabolites swinholide A and iso-swinholide A (Figures 1b–d).

Targeting these macrolides for isolation, fraction H was purified by C-18 solid phase extraction into subfractions H1–H7; these were analyzed using SMART 2.0 mixture analysis as well as MS\(^2\)-based molecular networking. The latter revealed a highly abundant and distinct cluster of features with the major m/z 1395.9 in active fraction H4 (Figure 1e). Isolation and structure elucidation of this compound led to the discovery of a new cytotoxic macrolide \(1\) (Scheme 1), named symplocolide A (1), representing a structural chimera between swinholide A (4) and luminaolide B (3).\(^\text{43}\) Analysis of the MS\(^2\) spectra of metabolites from these fractions using GNPS led to annotation of swinholide A (4), the glycosylated samholides A–I (5–13, Scheme 1 Figures S12–S22), as well as several putatively new analogues swinholides L and M, samholides J–M, and symplocolides B–D (Figures S23–S30).

Similarities in the MS\(^2\) fragmentation patterns of 1 and 4 (Figures 1f and g)\(^\text{29}\) suggested that they differed in their carbon backbones (see blue part of the feature m/z 939.49, Figures 1g, S4). This hypothesis was confirmed following isolation and unambiguous structure determination by 2D NMR analysis and HRMS (Figures S5–S11, Table S1). These analyses disclosed 1 to possess the backbone skeleton of 3 and the side chain structure of 4, thus constituting a chimera of these two macrolides.

All known swinholide-type compounds, irrespective of origin (sponge, cyanobacteria, algae, nudibranch), possess a highly analogous monomeric carbon skeleton and identical configurations at comparable chiral centers (Scheme 1); this has been confirmed in two cases by X-ray crystallography\(^\text{85,86}\) as well as total synthesis.\(^\text{87}\) These swinholide-type metabolites are produced by highly similar trans AT polyketide synthase biosynthetic gene clusters.\(^\text{29,45}\) Thus, the stereochemistry of 1 is assigned by analogy given the close similarities between the \(^1\)H and \(^13\)C NMR chemical shifts and coupling constants of compound 1 with those of previously reported swinholide-like macrolides (Tables S2, S3, Scheme S1).\(^\text{42–44,48}\)

Recently, Duggan et al. developed an “atomic sort” method wherein HSQC spectra from two different NP extracts were compared with a database of 1207, mostly human primary metabolites using Euclidian distance scores, and the method was able to detect novel structural features of marine NPs.\(^\text{39}\) This approach yielded a new analogue of a known antiprotozoal polyketide (gracilioether L, see SMART results in Figure S32).\(^\text{50}\) They also demonstrated the method with reisolation of known antibacterial cyclic heptapeptide, cyclo- marin A.\(^\text{51–53}\) However, three of four “novel features” detected by this method in cyclomarin A were false positives in terms of structural novelty (e.g., regular methyl group, methylene next to primary alcohol and alpha-carbon methine of N-methyl-leucine).

To demonstrate SMART 2.0 for mixture analysis and early prioritization for rare structural motifs, we used the NMR table wherein HSQC spectra from two different NP extracts were compared with a database of 1207, mostly human primary metabolites using Euclidian distance scores, and the method was able to detect novel structural features of marine NPs.\(^\text{39}\) This approach yielded a new analogue of a known antiprotozoal polyketide (gracilioether L, see SMART results in Figure S32).\(^\text{50}\) They also demonstrated the method with reisolation of known antibacterial cyclic heptapeptide, cyclo- marin A.\(^\text{51–53}\) However, three of four “novel features” detected by this method in cyclomarin A were false positives in terms of structural novelty (e.g., regular methyl group, methylene next to primary alcohol and alpha-carbon methine of N-methyl-leucine).

To demonstrate SMART 2.0 for mixture analysis and early prioritization for rare structural motifs, we used the NMR table from the SI in Duggan et al.\(^\text{49}\) to evaluate the HSQC data of the crude extract as well as pure cyclomarin A (not in the SMART database). Intriguingly, the top two hits for cyclomarin A and top 20 hits for the crude extract, consisting of 289 C–H correlations in the HSQC spectrum, revealed the closely related ilamycins C1 and C2 as candidate structures.\(^\text{34}\) Ilamycins harbor a very similar and rare \(\text{N}-(1,1\text{-dimethyl-2,3-epoxypropyl})\)-tryptophan moiety as cyclomarin A [substructure search of NP Atlas database (https://www.npatlas.org/)]
returned no additional NPs with this amino acid moiety].37 Furthermore, ilamycins contain a similar 4-hexenoic acid residue as cyclomarin A (Figures 2, S3). This example underscores that while features such as heteroatoms and
orthogonal information derived from genome mining of the aspects of NP drug discovery, especially when combined with continued development of SMART 2.0 can revolutionize features from crude extracts and fractions. We envision that the demonstrated here that SMART 2.0 can detect rare structural moieties, and is a great improvement over the weeks to favorably with the time-scale of LC-MS measurements and structure prediction takes approximately 8 s). This compares and structure prediction, took less than 30 min (the SMART by SMART, from NMR data acquisition to table construction visible in the HSQC spectrum. Finally, each analysis performed predicts structures with chemical moieties that are not directly visible peaks for interpretation. This aspect likely explains the positions of correlations. Thus, CNNs can extract information via analyzing chemical shift patterns as well as the relative HSQC experiment, CNNs can detect these features indirectly through analyzing chemical shift patterns as well as the relative positions of correlations. Thus, CNNs can extract information from nonpeak areas whereas spectroscopists mainly rely on visible peaks for interpretation. This aspect likely explains the extraordinary robustness of SMART 2.0, as it very accurately predicts structures with chemical moieties that are not directly visible in the HSQC spectrum. Finally, each analysis performed by SMART, from NMR data acquisition to table construction and structure prediction, took less than 30 min (the SMART structure prediction takes approximately 8 s). This compares favorably with the time-scale of LC-MS measurements and annotations, and is a great improvement over the weeks to months typically required for the structural analysis of highly modified cyclic peptides such as cyclomarin A.

In this report we introduce a highly improved user-friendly version of the SMART tool and demonstrate its remarkable ability to rapidly recognize natural product structure types. SMART 2.0 is now available online to the academic community (https://smart.ucsd.edu/classic).

Rapid structure prediction of major constituents from crude extracts and fractions greatly assists with prioritization of structurally novel or otherwise interesting NPs for further study. Because knowledge of the structural class of an unknown NP can greatly accelerate its structure elucidation, this tool can overcome several of the bottlenecks currently present in NP research. We applied SMART 2.0 on crude fractions from a cyanobacterial extract that helped target the isolation of a new chimeric macrolide, symlocilode A, as well as dereplicate swinholide A and related compounds. The hybridic nature of symlocilode A suggests a novel biosynthetic pathway, an aspect that is under continuing study as the genetic basis for molecular diversification in a compound class is of great interest.43,55 Further, we demonstrated here that SMART 2.0 can detect rare structural features from crude extracts and fractions. We envision that the continued development of SMART 2.0 can revolutionize aspects of NP drug discovery, especially when combined with orthogonal information derived from genome mining of the biosynthetic gene clusters, accurate mass and fragmentation data, and fingerprint analyses of molecular substructures via the GNPS platform, MS2LDA, and Sirius.22,27,28

Figure 2. Comparison of the “atomic sort” method and SMART 2.0 to detect novel/rare structural elements. (a) HSQC correlations suggesting structural novelty by “atomic sort” method (highlighted in blue). (b) Top two results of SMART 2.0 analysis querying cyclomarin A (experimental HSQC data from reference49) to detect related compounds that include rare structural moieties (highlighted in red, purple).
Author Contributions


Notes

The authors declare the following competing financial interest(s): Chen Zhang, Garrison W. Cottrell, and William H. Gerwick are the cofounders of NMR Finder LLC. Mingzun Wang is the founder of Omega Laboratories LLC.

ACKNOWLEDGMENTS

The authors are grateful for funding from NIH grant R01 GM107550 to G.W.C., P.C.D., and W.H.G. and Gordon and Betty Moore Foundation grant GBMF7622 to G.W.C., P.C.D., and W.H.G. We thank N. Moss and B. Miller for assistance with collection of the Symploma sp. and G. Arevalo for extraction and VLC fractionation. We also thank P. Landon for purchasing and assembling our Graphic Processing Unit (GPU) machine.

REFERENCES


