IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

SMALL MOLECULE ACCURATE RECOGNITION TECHNOLOGY(SMART) TO ENHANCE NATURAL PRODUCTS RESEARCH

[1]MAGGIDI RAMANI, [2]GOVARDHAN SREENANCHA, [3]SRIKANTH NYAMTABAD
[1,2,3]Assistant Professor
Department of Chemistry
Kshatriya College of Engineering

ABSTRACT:

Various algorithms comparing 2D NMR spectra have been explored for their ability to dereplicate natural products as well as determine molecular structures. However, spectroscopic artefacts, solvent effects, and the interactive effect of functional group(s) on chemical shifts combine to hinder their effectiveness. Here, we leveraged Non-Uniform Sampling (NUS) 2D NMR techniques and deep Convolutional Neural Networks (CNNs) to create a tool, SMART, that can assist in natural products discovery efforts. First, an NUS heteronuclear single quantum coherence (HSQC) NMR pulse sequence was adapted to a state-of-the-art nuclear magnetic resonance (NMR) instrument, and data reconstruction methods were optimized, and second, a deep CNN with contrastive loss was trained on a database containing over 2,054 HSQC spectra as the training set. To demonstrate the utility of SMART, several newly isolated compounds were automatically located with their known analogues in the embedded clustering space, thereby streamlining the discovery pipeline for new natural products.

INTRODUCTION:

As a discipline, natural products research (NPR) enables and benefits numerous downstream research fields, such as chemical biology, chemical ecology, drug discovery and development, pharmacology and the total chemical synthesis of natural products (NPs). In this regard, approximately 70% of all approved drugs are NPs, their analogues, or a chemical modification of an existing NP1 . In addition to these academic and societal benefits, NPR provides a powerful incentive for the conservation and sustainable use of biodiversity and biodiverse habitats.

An important step in NPR is dereplication, the process of assessing the uniqueness of a new compound in relationship to all known ones. In most NPR, traditional compound dereplication practices have entailed the collection and analysis of nuclear magnetic resonance (NMR) spectra, including running 1D and 2D NMR spectroscopic experiments for the purposes of molecular framework construction, assembly, and relative stereochemistry determination. More recently, mass spectrometric approaches and mass spectrometry (MS)-based molecular networking3 , in part stimulated by integration with DNA sequencing and genome mining4,5 have been integrated into NPR workflows. Nevertheless, conventional NMR practices are still indispensable to the characterization and dereplication of NPs. Unfortunately, 2D NMR experiments can be time consuming, especially when the sample is relatively scarce. Furthermore, 2D NMR-based structural assignments can sometimes take protracted periods of time to accomplish due to the inherent structural complexity of some NPs.

Along with relatively recent improvements in mass spectrometry, circular dichroism and infrared spectroscopy techniques, state-of-the-art cryoprobe NMR instruments have reduced the sample requirements for NPs discovery to just a few nanomoles6 . Nevertheless, acquisition of NMR spectra may still require a relatively large number of time consuming scans before Fourier transformation of the data. Furthermore, conventional 2D NMR spectroscopy relies upon linear sampling of the frequency evolution in the indirect dimension (usually the 13C dimension). When generating high frequency resolution in the indirect dimension, extensive sampling is required and the experiments become very time consuming. Modification of conventional uniform sampling to non-uniform sampling (NUS)7–13 allows the number of experiments in the indirect dimension to be reduced, thereby reducing the overall time of the experiment. The NUS method is designed to reduce the number of data collection experiments while at the same time delivering an accurate estimation of the fully sampled spectrum.

To streamline compound dereplication or even structure determination, algorithms have been applied for 2D NMR spectra comparisons, such as the 2D NMR peak alignment algorithm14,15. However, these techniques are not powerful enough to accurately classify 2D NMR spectra into the correct NP family. This arises for several reasons, such as compound concentration, solvent effects, and the interactive effect of a single functional group alteration on 1 H and 13C NMR chemical shifts, all of which combine to increase the difficulty for computer assisted 2D NMR data analysis. At the same time, artefacts are often introduced into NMR spectra, and this makes it difficult for existing pattern recognition or overlap methods to distinguish genuine peaks from artefacts. However, artificial intelligence technologies, such as deep learning16,17, have generated new approaches for meeting these challenges. Compared with conventional machine learning methods, which require the cumbersome process of selecting and creating features that might be suboptimal for a given task, deep learning allows creation of the most suitable set of features within the process of training, without any design or involvement by the investigator. Moreover, some deep learning methods work well even when the number of categories is very large and unknown during the training process. Thus, deep learning is an ideal method by which to analyse and categorize 2D NMR spectra of NPs. For NPs, there are an essentially unlimited number of categories for different compound families, with many being unknown at the present time. Additionally, it is quite common for each category to contain fewer than 50 different members; in the work of our laboratory with marine cyanobacterial NPs, this is the case for all of the molecular families we have encountered over the past 40 years, including the curacins18–20, apratoxins21, lyngbyabellins22 and majusculamides.

Other approaches for automatic recognition of NMR spectra have appeared in the literature or private sector. The typical approach is to create grids over the data and then compute similarities based on how many points fall into the same grid cells26. This approach can miss peaks that are near one another that happen to fall in different grid cells, so an enhancement of this approach is to use multiple grid resolutions and offsets before computing the similarities27. Our convolutional network approach automatically does this by using overlapping convolutions

combined with increasing-sized receptive fields through pooling the results from previous layers. However, our method of deciding similarity is learned by the network through nonlinear dimensionality reduction via training it to map together those compounds it recognizes as being from the same family, and to map different families to different locations in the underlying space.

Another method involves computer-aided structure elucidation (CASE, ACD/Labs) which is largely based on applying a least-squares regression (LSR) approach for comparing NMR chemical shifts; this tactic is not powerful enough to satisfactorily accommodate solvent effects, instrumental artefacts, or weak signal issues14,15. An early effort using machine learning applied to NMR spectra was reported in (Wolfram et al., 2006)28. They used Probabilistic Latent Semantic Indexing (PLSI), a method usually applied to text documents for information retrieval purposes. PLSI maps documents into a lower dimensional space using a probabilistic analogue to Singular Value Decomposition (SVD) applied to a document by word count matrix. To apply PLSI to compounds, the typical multi-scale and shifted grid cell approach was used, treating each grid cell as a "word" in the compound. This is essentially learning a linear mapping from the feature space to a reduced space, and thus is not as powerful as using a nonlinear deep network.

In our approach, heteronuclear single quantum correlation (HSQC)29 spectra are recorded using a 2D NMR pulse sequence that uses the large heteronuclear coupling between directly bonded nuclei within an organic molecule to correlate directly bonded atoms (e.g. 1 H and 13C, with 1 H being defined as the direct dimension and 13C the indirect dimension). The peaks of those correlated nuclei in the 2D HSQC spectra are generated by detecting coherences that connect states whose total z-angular momentum quantum numbers differ by one order (i.e. single-quantum coherences). In this regard, an HSQC spectrum is deemed as the 'fingerprint' or 'face' for a natural product molecule, and thus is highly discriminating. Specifically, within a 2D HSQC spectrum, signals in the direct dimension can be distinguished if they have shifts of 0.02 ppm or greater, and in the indirect dimension if they have shifts of 0.1 ppm or greater. Furthermore, most 1 H chemical shifts occur between 0.5 and 9.5 ppm, whereas in the 13C dimension chemical shifts typically occur between 10 and 215 ppm, which gives rise to 922,500 distinguishable positions within a 2D HSQC spectrum. When summed over all protonated carbons in a molecule of 20 carbons with attached protons, the number of potential combinations is in the tens of millions, and is thus highly discriminatory. In addition, this technique avoids detection of double-quantum coherence, resulting in relatively few artefacts. In contrast, the commonly used heteronuclear multiple bond correlation (HMBC) experiment detects two and three bond correlations by selecting smaller multiple bond heteronuclear coupling constants (around 5–10Hz for 1 H-13C versus one bond of 125–170Hz) for double-quantum and zero-quantum transfer. Therefore, while the HMBC experiment produces an even larger amount of theoretical information, it is prone to introducing artefacts and its complexity makes it more difficult to interpret. In addition, the HSQC when performed with NUS discussed above is a relatively quick and efficient experiment for data accumulation.

Here, we report the development of the Small Molecule Accurate Recognition

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

Technology (SMART) prototype, a system that integrates the benefits of NUS NMR with advances in deep learning to enhance and improve the efficiency of NP dereplication. To create SMART, a database of training examples containing 2D HSQC spectra of 2,054 compounds was compiled. These examples were used to train a deep network that learns to map the spectra into a cluster space where similar compounds are near one another and dissimilar compounds are far apart. To perform this function, we use a deep convolutional neural network (CNN) employing a siamese architecture30 as described in the methods section. A siamese network amplifies the number of training examples by training on pairs of spectra that are labelled "same" or "different," rather than training on individual examples. The network then learns features of the spectra that are relevant to their similarities and differences, and uses this to create the cluster space. The resulting mapping then generalizes to new compounds, placing them in the space near compounds with similar HSQC spectra. We evaluate SMART by holding back several known NPs from different families from the training set, and then show that SMART maps them into their proper location within the cluster space. We also present here the rapid identification of a newly isolated natural product compound family as a result of SMART's ability to cluster similar compounds together. HSQC spectra were collected for several nonribosomal peptide synthetase (NRPSs)-derived NPs that had been isolated from two marine cyanobacteria. These novel spectra were accurately mapped by SMART into the 'viequeamide' subfamily of NPs.

RESULTS AND DISCUSSION

The SMART prototype. SMART is a user-friendly, AI-based dereplication and analysis tool that uses 2D NMR data to rapidly associate newly isolated NPs with their known analogues. SMART has been designed to mimic the normal path of experiential learning in that additional 2D NMR spectral inputs can be used to enrich its database and improve its performance. In short, SMART aims to become an experienced associate to natural products researchers as well as other classes of organic chemists. The SMART workflow consists of three steps, 1) 2D NMR data acquisition by NUS HSQC pulse sequence, 2) 2D NMR spectral analysis by deep CNN, resulting in an embedding of the spectra into a similarity space of NPs, and 3) molecular structure dereplication or determination by the user (Fig. 1). This process gives users rapid access to a well-organized map of structurally determined NPs, and helps ensure that SMART's insights are chemically rational. In this regard, SMART capitalizes on the wealth of molecular fingerprints, namely 2D HSQC spectra, built over the past four decades31,32, and reciprocally, we anticipate that 2D HSQC spectral databases will experience an accelerating expansion as a result of SMART's application.

The workflow (Fig. 1) of SMART begins with recording the NUS HSQC spectrum for a pure small organic molecule; in the case of NPR, this is a substance extracted and purified from an organism of interest, but just as easily could be a small molecule produced from organic synthesis, biosynthesis or from a metabolomic study. A small molecule is defined here as one whose transverse relaxation time constant (T2) is on the same order of magnitude as its longitudinal relaxation time constant (T1) when dissolved in liquid solution. In other words, the nuclear spins of a small

molecule should be synchronized between 107 to 108 Larmor precession cycles during a liquid state 2D HSQC experiment33. Nevertheless, the SMART concept is not inherently confined to small molecule NUS NMR spectra, considering the ability of NMR to structurally characterize molecules of many sizes and types. NUS HSQC experiments are highly advantageous for small molecule structure elucidation compared with conventional pulse sequences due to their rapid acquisition, few spectral artefacts, and intrinsic high resolution. Nevertheless, as discussed below, conventional 2D HSQC spectra can be provided to the AI system and spectral recognition achieved. In fact, the initial database of HSQC spectra that were compiled to train the SMART system was acquired in this manner.

Due to lower sampling density, NUS HSQC requires alternative approaches to convert the indirectly sampled time domain into visual spectra of the frequency domain, and thus methods other than the Discrete Fourier Transform are required. To this end, Iterated Soft Thresholding (IST)34,35 followed by the Maximum Entropy Method (MEM)36,37 was applied to NUS data collected for the model compound strychnine. In order to achieve convergence to a local minimum, a Lagrange multiplier was applied to weight the regularization function, the L1 norm, in the IST routine. Previous studies12 have shown that IST is superior to Maximum Entropy Reconstruction (MaxEnt)38 (not to be confused with MEM) in NUS NMR data reconstruction, owing to the simplicity of IST with fewer adjustable parameters and the resultant ease of application. Nevertheless, IST suffers slower convergence compared to MaxEnt for spectra with a high dynamic range. However, it has been shown that changing the step sizes in IST can achieve visualization of the final spectra indistinguishable from those reconstructed by a well-tuned MaxEnt process39. The MEM can then be applied after Fourier Transformation of the IST reconstructed data in the direct dimension, resulting in an improvement that derives from the fact that MEM is biased towards the enhancement of smaller line widths40. For the model compound, the HSQC correlation signals of the C-11 methylene protons (3.11 ppm and 2.67 ppm) to their subtending carbon were visibly strengthened after sequentially applying IST (400 iterations) and MEM (3 iterations) compared with application of IST (400 iterations) with Linear Predictions (LP) during data reconstruction of the non-uniformly sampled 2D NMR spectra (Fig. 2).

g. 2). Our deep learning method is based on a siamese neural network architecture41. A siamese network is comprised of a pair of identical networks that are trained with pairs of inputs. These are mapped to a representational space where similar items are near one another and different items are further apart. As a result, it produces a clustering of the input space based on a similarity signal. In our case, it first maps the input HSQC spectra into a ten dimensional space, which then can be mapped into a two dimensional space by Principal Components Analysis (PCA) for visualization purposes

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

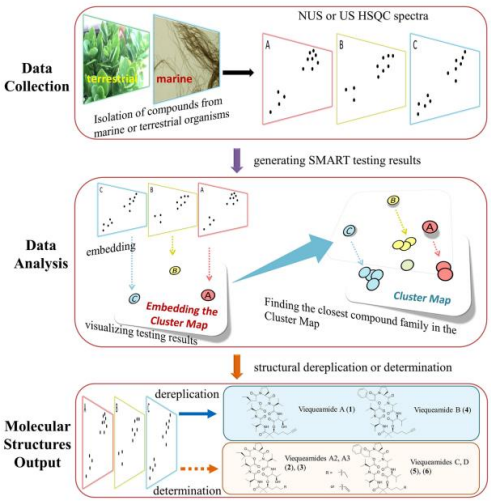

Figure 1. Workflow for the Small Molecule Accurate Recognition Technology (SMART). Experimental HSQC spectra of newly isolated pure natural product molecules collected using either NUS HSQC pulse sequences or conventional HSQC techniques, are automatically embedded by SMART into a cluster space near similar, previously-characterized compounds. The resultant embedding in the cluster map is visualized using the Bokeh visualization package72, where each node represents an HSQC spectrum processed by SMART. The node colours in a local area of the clustering map designate compounds from the same journal articles and thus of the same natural product family. This facile method allowed tracking of compounds into SMART, but is not of paramount significance in that some compounds reported in different publications display closer relationships in SMART and by structural comparison than to compounds within the same article. When available, the node labels are the compound names; otherwise, the labels are for the organism from which the compound derives. Node distance is proportional to relatedness, a quantification of molecular structural similarity. The 2D cluster map is created by performing Principal Component Analysis (PCA) of the 10D space outputs to reduce to 2D. Optionally, the top 5, 10 and 20 closest nodes in the 10D space are available in text format. The proof-of-concept experiments are illustrated: Dereplication (solid blue arrow) of viequeamides A (1) and B (4), and determination (dashed orange arrow) of viequeamides A2 (2), A3 (3), C (5) and D (6), isolated from 1) Rivularia sp., collected in Vieques, Puerto Rico and 2) Moorea sp., collected in American Samoa.

are trained by backpropagation of errors45. CNNs are structured to learn local visual features that are replicated across the input, hence the term “convolutional”. The local maximum of these features are then input to another layer that learns local features over the previous layer of features, and this process is repeated for several layers. In previous work, it has been shown that the feature maps resulting from each convolutional layer become more abstract as the layers of the network are traversed. We show the first layer features in Fig. 3. By using the local maxima of feature responses over nearby locations in the input, the network will generalize to patterns that are shifted in the (, f f) 1 2 plane of the spectra, i.e., it achieves some translation invariance. Thus, the network is inherently hierarchical, like the mammalian visual system, and learns more and more abstract features in deeper layers of the network. In a siamese network, the final layer is not trained to classify the inputs; instead, a set of units are trained to give similar patterns of activation for similar inputs (as given in the teaching signal) and different patterns of activation for inputs that are labelled as different. Hence, they produce a clustering in the space of unit activations46.

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

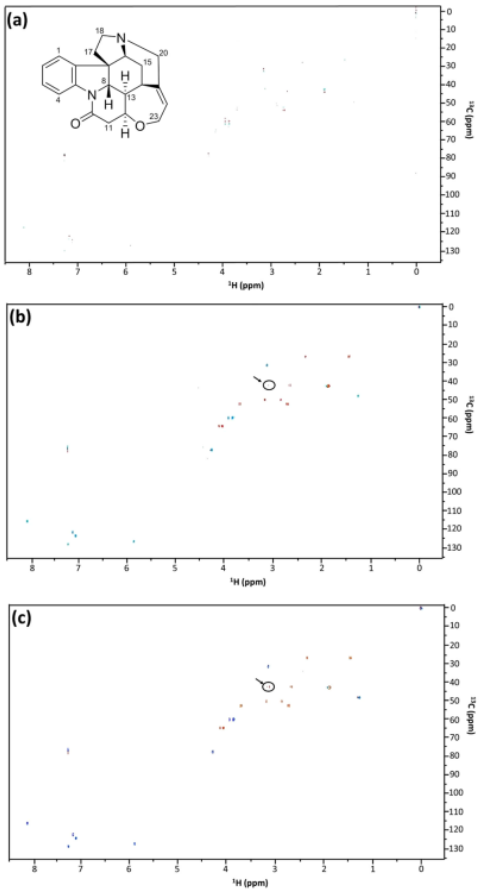

Figure 2. Data reconstruction results of a non-uniformly sampled (NUS) HSQC experiment. All of the three full HSQC spectra were recorded with a 50 nmole strychnine sample in CDCl3 on a 600MHz Bruker 1.7 mm cryoprobe instrument, using 32 out of a total 128 increments (25% sampling density) in the indirect dimension and 8 scans. The differences between the three spectra were that (a) was transformed with the maximum entropy method (MEM) alone, (b) was transformed with the iterative soft thresholding (IST) alone, and (c) was transformed with IST followed by MEM. The doublet (see black arrows and circles in (b) and (c)) associates with the protons on the methylene (C-11) adjacent to the ketone in strychnine (see text for discussion).

As a result, molecules that are similar in HSQC spectra will be mapped to nearby locations in the output space. If the network generalizes well, it will place novel molecules near known ones that have similar NMR spectra. This allows the system to rapidly identify candidate known molecules that may have similar chemical features to the novel molecule, allowing the user to search through a small subset of known molecules for similar compounds. In our initial approach, we used ten output units (i.e., a 10 dimensional space), which can be visualized by applying Principal Components Analysis (PCA) to reduce the 10 dimensions to two.

Network training and results. The neural network was trained using stochastic gradient descent47 with the Adagrad48 update rule. To speed the training, we employed batch normalization49, which reduces the internal covariance shift by standardizing the distribution of features on each layer. The network was found to train 7 times faster (wall clock time) using batch normalization. When training the CNN, the datasets (see the Methods section for details) were divided into three subsets; the training set containing 80% of the spectra, used to adjust the parameters of the network, the validation set

Figure 3. Features learnt by the first convolutional layer of the CNN. Feature maps were extracted from convolution layer 1 in Table 1, with the eight blocks of

4×4 pixels in this figure corresponding to the results of each of the eight filters applied to the HSQC dataset.

Layer Number	Layer Type	Number of Filters (Stride 1)	Size	Additional Information
1.	convolutional	8	4 × 4	maxpool 4 × 4 stride 2
2.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
3.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
4.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
5.	fully connected	—	128	dropout 0.5
6.	fully connected	—	128	dropout 0.5
7.	fully connected	—	128	dropout 0.5
8.	fully connected	—	*K*	*K*-dimensional embedding layer

Table 1. The Architecture of the Deep CNN Used in This Studya . a The dimensionality of the input data is 512×512.

containing 10% of the data used for early stopping, and a test set containing the remaining 10% of the data (for details, see Methods). The test set consisted of HSQC spectra that were not used during the training process. The error from the validation set was monitored to prevent overfitting on unseen data. The test spectra were then embedded in the cluster map to locate their nearest neighbours. In this way, the test HSQC spectra were matched with other structurally similar compounds (e.g., from the same compound family or by possessing a high Tanimoto similarity score).

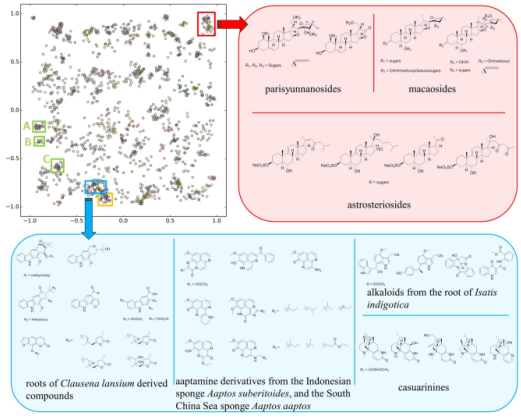

Figure 4. The SMART cluster map based on training result of 2,054 HSQC spectra over 83,000 iterations, with inset boxes representing different compound classes discussed in the text.

training iterations (Figures S1 and S2 in the Supplementary Information for the cluster map with analysis), and subsequently, we trained on a larger dataset of 2,054 for a total of 83,000 iterations. The tight structural similarity between the compounds and their locations in the cluster map is evident (Fig. 4).

Related work. Again, the aforementioned grid-cell approaches28 are similar to ours in that the shifted grid positions can be thought of as corresponding to the first layer of convolutions, which have small receptive fields (like grid cells), and they are shifted across the input space like shifted grids. Also, our approach uses layers of convolutions that can capture multi-scale similarities. The grid-cell approaches, however, use hand-designed features (i.e. counts of peaks within each grid cell), and the similarities are computed by simple distance measures. In particular, PLSI and LSR are linear techniques applied to hand-designed features. Furthermore, other representations, for example the 'tree-based' method59, also rely on data structures designed by the researcher. Our approach, using deep networks and gradient descent, allows higher-level and nonlinear features to be learned in the service of the task. This approach is similar to modern approaches for computer vision, which since 2012 has shifted away from hand-designed features to deep networks and learned features, and has led to orders of magnitude better performance. Similarly to how deep networks applied to computer vision tasks have learned to deal with common problems, such as recognizing objects and faces in different lighting conditions and poses, our CNN

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

pattern recognition-based method can overcome solvent effects, instrumental artefacts, and weak signal issues.

SMART recognition of noisy HSQC spectra. Because white Gaussian noise is often seen in experimental HSQC spectra, we investigated the robustness of the SMART to recognize HSQC spectra in the presence of significant noise. By adding noise to HSQC spectra in the SMART10 database and measuring the Euclidean distance of those noisy spectra to their original ones, we were able to observe that as noise intensity increases so does the distance increase from the original location in the 2D cluster map. However, the noisy spectra were still effectively recognized as being closely related to their original compounds (Fig. 7 and Supplementary Information).

SMART characterization of Viequeamides of NRPS origin. As a practical example of the functional use of the SMART workflow to discover new NPs, we used it to rapidly characterize a group of unknown marine cyclic depsipeptides from two marine cyanobacteria: 1) Rivularia sp. collected in Vieques, Puerto Rico and 2) Moorea sp. collected in American Samoa. These compounds were isolated in the course of our ongoing efforts to discover marine cyanobacterial NPs with anti-cancer properties60. Metabolites from these two collections were purified by high performance liquid chromatography (HPLC), and then 1 H-13C HSQC data were collected with 100% sampling density, but using the NUS pulse sequence in the indirect dimension for all six purified compounds. Data reconstruction as described above for the six samples yielded HSQC spectra, and these were subjected to the SMART workflow to embed them in the cluster map. We found that the six nodes clustered with nodes for the previously characterized viequeamides A (1) and viequeamides B (4). After an analysis of various 2D NMR spectra, and MS, IR and UV data, the planar structures of the four new compounds were determined (Fig. 1, compounds 2, 3, 5, 6). The absolute configurations of these compounds were then elucidated by Marfey's analysis and/or X-ray crystallography, completing their structure determination. Evaluation of the toxicity of the pure compounds to H-460 human lung cancer cells revealed that two possessed relatively potent cancer cell toxicity properties; viequeamide A2 (2) had an IC50=0.62 μM and viequeamide A3 (3) had an IC50=1.98 μM. Viequeamides B (4), C (5) and D (6) showed no appreciable H-460 human lung cancer cytotoxicity.

METHODS

Training set collection and processing. The dataset of HSQC spectra was compiled from available online sources. We removed spectra that showed no peaks (i.e., the spectra in the publication appeared blank). We collected all usable1 H-13C HSQC spectra (4,105 in total), including a few cases of the same compound in different deuterated solvents, from the supporting information of the Journal of Natural Products, years 2011, 2012, 2013, 2014 and 2015. In addition, the HSQC spectra of somocystinamide A61 and swinholide A62 in the supporting information of Organic Letters were also included in the dataset. Around 2,056 spectra were removed from this series, because their molecular class had less than 5 HSQC spectra. All spectra were collected and initially

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

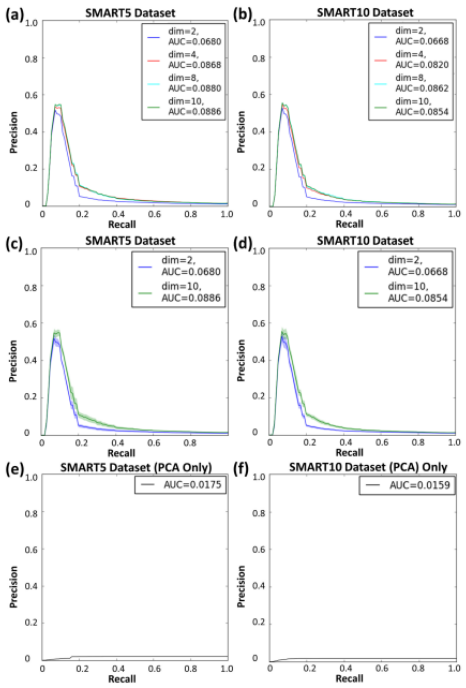

Figure 5. Precision-recall curves measured across 10-fold validation for different dimensions (dim) of embeddings. (a) and (b) Mean precision-recall curves on test HSQC spectra for SMART5 and SMART10 datasets, respectively. (c) and (d) Mean precision-recall with error curves (grey) for SMART5 and SMART10, respectively. (e) and (f) Mean precision-recall curves for SMART5 and SMART10 clustered by Principal Component Analysis (PCA) without use of the CNN. AUC: area under the curve.

processed by the following steps: (1) The HSQC spectra were saved as png format grayscale images at a resolution of 512×512 pixels (the minimum resolution in the proton dimension is 51.2 pixels per ppm and in the 13C dimension it is 2.8 pixels per ppm.); (2) lines surrounding spectral edges, annotations, chemical structures, and other annotations were deleted using Photoshop such that only the HSQC signals and noise were present in the

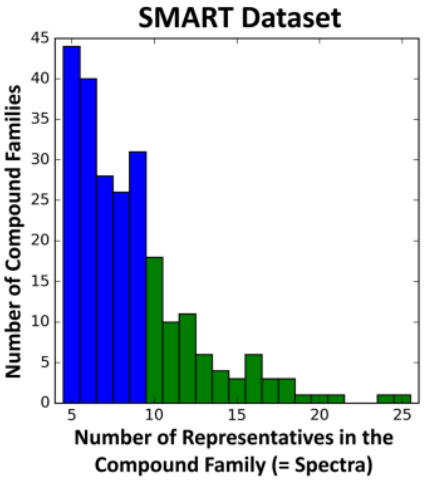

Figure 6. Distribution in the Training Dataset of Numbers of Families Containing Different Numbers of Individual Compounds. The SMART5 training set contains 238 compound subfamilies, giving rise to 2,054 HSQC spectra in total. (Blue and Green) The SMART10 training set contains 69 compound subfamilies and is composed of 911 HSQC spectra in total. (Green only).

images; (3) images were rotated and/or flipped when necessary to ensure that the horizontal dimension was the direct1 H dimension with chemical shifts increasing from right to left, and the vertical dimension was the indirect13C dimension with chemical shifts increasing from top to bottom; (4) images were uniformly converted into black (signal and noise) and white (spectral background); (5) images from the same publication were pooled and labelled as the same training class, as all of the publications we used reported compounds from a single family; (6) a cross shaped 3×3 median filter63 was applied to remove unwanted salt-and-pepper noise; however, no other enhancements were applied (Figure S4 in the Supplementary Information for an example of spectra input preparation). Essentially, in this study, the relevant quantity for measuring similarity was the positions and shapes of the various peaks

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

relative to one another, rather than their absolute positions

NUS 2D NMR data generation. In order to generate an independent test set, we developed an optimized NUS pulse sequence using an NMR standard (strychnine, 50 nmole TCI America, Catalog No. S0249). This optimized method was then applied to several newly isolated NPs (e.g., the viequeamides). The viequeamides were isolated from two different marine cyanobacteria; Rivularia sp. collected in Vieques, Puerto Rico60 and Moorea sp. collected in American Samoa. Detailed isolation and structural elucidation of these compounds will be published separately. The 2D NMR spectra were recorded on a 600MHz Bruker Avance III spectrometer with a 1.7 mm Bruker TXI MicroCryoProbeTM using TopSpin 2.1. The solvent CDCl3 contained 0.03% v/v trimethylsilane (δH 0.0 and δC 77.16 as internal standards using trimethylsilane and CDCl3, respectively). All spectra were recorded with the sample temperature at 298 °K.

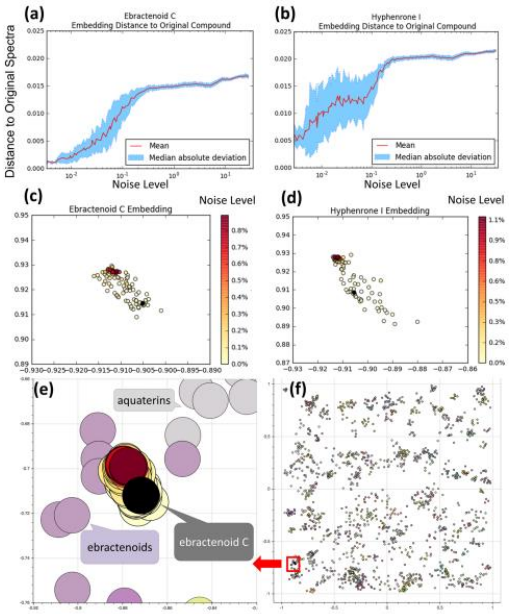

Figure 7. Distance of the noisy spectra measured against the original spectra of ebractenoid C and hyphenrone I. The distance measure in the y axis of the ebractenoid plot (a) and hyphenrone plot (b) is the same as the cluster map in Figs 4 and 7(f). The noise level is defined by dividing pixels altered over the total number of pixels of the HSQC spectra. The results visualized in the 2D cluster maps with each node representing one noisy spectra, and with node color intensity as a function of the noise level, for the ebractenoids (c) and hyphenrones (d). The original image without added noise is shown as the black node in these 2D cluster maps. We then embedded the nodes for the ebractenoids in (c) to a global view of the 2D cluster map in (f), and zoom in on the red box in (f) as shown in (e). Note, larger node sizes are used to depict compounds in (e) versus (c).

The deep siamese network. As depicted in Table 1, the overall deep CNN siamese architecture used in this study is similar to AlexNet42, and consists of 8 layers comprised of 4 convolutional layers followed by 4 fully connected layers. This network is used as the two "twins" in the siamese network. The output layer contains vectors in PK. Here, K is the embedding dimension. The energy loss function defined in equation 2 (below) is applied to the outputs of the embedding layer (layer 8). We ran several experiments to find the best K and measured the accuracy on the validation set. Empirically, for the given dataset, K = 10 gave us the best results.

Loss function.

Siamese networks are trained with an energy function that is minimized by gradient descent. The design of the energy function determines the way in which pairs of items are pushed together or pulled apart. There are at least two such functions that have been used30 in the literature; here, we used a modified version of the

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

spring model developed by Hadsell et al. 41. The energy function is described with the following notation; for example i, the input vector is represented as xi , and the output label as yi . The output label is defined as a "one hot" vector, where if there are k categories, yi is a k-dimensional binary vector, and if the category is c, yi is 1 at the cth position and 0 everywhere else. Meanwhile, xi , the input HSQC spectra, is treated as a vector. We treat our neural network as a function GW, where W is the weights of the network. Then the output of the neural network is GW( ) x . GW( ) x is a vector of dimension K, a hyperparameter of the system. We then define the distance function d between images xi and xj :

$$d(x_i, x_j) = \left\| G_W(x_i) - G_W(x_j) \right\| \quad (1)$$

where · is the Euclidean distance function. Now we can define the energy function L to be minimized as41:

$$L(x_i, x_j) = \begin{cases} \frac{1}{2} \max\left(0, d(x_i, x_j) - m\right)^2, \text{ if } y_i = y_j \\ \frac{1}{2} \max\left(0, m - d(x_i, x_j)\right)^2, \text{ otherwise} \end{cases} \quad (2)$$

Training details of the siamese network. We implemented our system using the Theano67 and Lasagne (http://tinyurl.com/hl9dy9y) deep network packages. The siamese network was trained using mini-batch stochastic gradient descent with the Adagrad45 update rule, following the protocol introduced by Hadsell et al. 41. Specifically, 50% of each mini-batch has negative samples ((x x , , y y,) i j i j s.t. ( ) y y) i j ≠ , and 50% has positive samples ((x x , , y y,) s t. .( ) y y =) i j i j i j . The Adagrad update rule tunes the step size automatically in real time, making learning stable in later iterations. We used hyperbolic tangent as the activation function for all layers including the output layer. The weights were initialized using Xavier initialization68. The initial learning rate was α = . 0 001, and the mini-batch size was 256. We applied dropout regularization69 on layers 5, 6, and 7 of the network, and batch normalization49. We found that applying batch normalization speeds convergence by a factor of 7. The total number of parameters in the network is 399,102, considering that the number of parameters triples when batch normalization is applied. We used Amazon EC2 instances to run our experiments.

We also used 10-fold cross validation to estimate performance (Figs 8 and 9). Specifically, a different 10% of the training set was held out as a test set 10 times, and the results were averaged to report performance. For each fold of the cross validation, we held out 10% of the data for early stopping. In this way, all of our HSQC spectra were used for testing. Here, the complete split was 8:1:1, training:validation:test. The iterations stop at the point in training where the error on the hold-out set is minimized. Here, the error was a measure of average precision on the hold-out set.

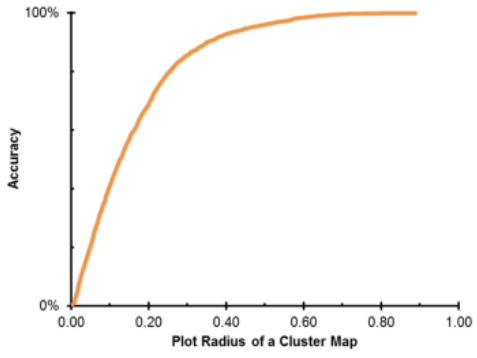

Figure 8. Plot of the Accuracy of SMART as the radius around a project point increases. This figure shows the fraction of correct families captured by a hypersphere of the given radius around each node in the cluster map. The distances between nodes in the cluster map has no physical meaning, but is a quantification of HSQC

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

spectral similarity. SMART can achieve good accuracy (proper placement in the map of a new compound to its correct compound family) within 0.5 radius of a 2-dimensional cluster map, and even better for a 10-dimensional map.

Validation of the model on "novel" categories. To evaluate whether the system performs properly with new categories of molecules, we performed the following three experiments. In SMART5, we removed the HSQC spectra of three categories of compounds (ebractenoids, naphthomycins, and veraguamides) from each of three common NP families (terpenoids, polyketides, and peptides, respectively), for each experiment, and used those removed spectra as a test set. During training, each subfamily was given a different label, however, this information was only provided to the training algorithm in the sense of "same/different category" as in Equation 2. This experiment thus tested whether a subfamily of terpenoids that was unfamiliar to the network would be mapped close to the other terpenoids. For example, there are 10 compounds in the terpenoid subfamily of ebractenoids that were not used during training. During testing, they were presented to the network, and their distance to the other terpenoids measured. This experiment was repeated for the naphthomycins, and veraguamides, and their location in the embedding space was evaluated for whether they were properly mapped to their respective families (e.g. polyketides and peptides, respectively). This experiment revealed that the ebractenoids clustered with the terpenes and terpenoids in the 10-dimensional space (Table S2). Similarly, the naphthomycins and veraguamides were subjected to a similar experiment (Table S3,S4) and confirmed that SMART was able to properly place compounds to which it was naïve.

Recognition of noisy HSQC spectra. Using Matlab 2013, we created a 2D matrix of white Gaussian noise to simulate the noise in real-time measurements. Next, we applied 2D Fast Fourier Transform (FFT) to this 2D noise matrix. The transformed FFT results for these noisy spectra were sized to match those of two randomly selected compounds (hyphenrone I and ebractenoid C) from the SMART10 database57,71. We also calculated the noise intensity in the spectra by dividing the number of noisy pixels by the total number of pixels. The noise matrix was then added to the two HSQC spectra, and the intensity of the noise was then increased consecutively in a finite arithmetic progression of 140 steps, rendering 140 noisy spectra for each compound. In addition, at each noise level, the white noise was again randomized 100 times, rendering a total of 14,000 noisy spectra. These noisy HSQC data were then processed by the convolutional neural networks pre-trained with SMART10 for over 10,000 iterations. The results are shown as two distance vs. noise plots in Fig. 7(a) and (b). The distance measure displayed in the vertical axis of these two plots was in the same units as the cluster map in Fig. 4. The results were also visualized in 2D cluster maps with each node representing one noisy spectrum, with the intensity of the node color representing the noise level (Fig. 7(c) and (d)). The original image without added noise is shown as the black node in those 2D cluster maps. In order to further visualize the internode distance between nodes that represent noisy spectra and those that represent our training dataset, we embedded the nodes of the noisy spectra

ISSN PRINT 2319 1775 Online 2320 7876
Journal Volume 10, Iss 9,2021

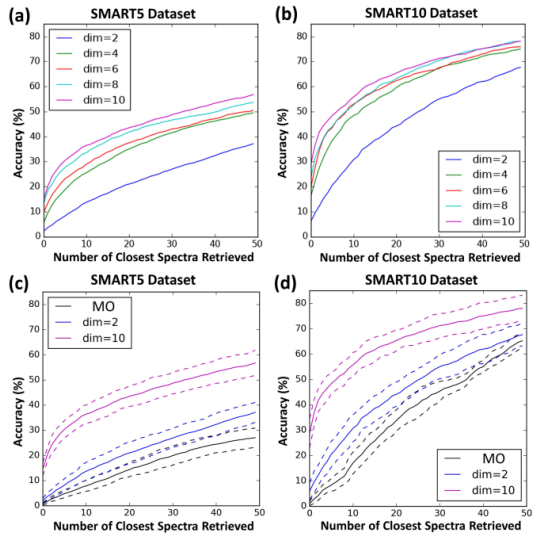

Figure 9. Closest retrieval curves measured across 10-fold validation for different dimensions (dim) of embeddings. For (a) and (b), mean closest retrieval curves on test sets for SMART5 and SMART10 datasets, respectively. For (c) and (d), mean closest retrieval curves with error curves ($\mu \pm \sigma$, dashed lines) for SMART5 and SMART10, respectively. In (c) and (d), the black plot (MO, most frequently occurring) is a baseline prediction of random compound associations on the basis of the number of members in a compound family. Specifically, the category with the most members is picked as the first compound association, the second most members as the second one, etc. This order is the same irrespective of the compound being considered.

in Fig. 7(c) in a global view of the 2D cluster map shown in Fig. 7(f), and provided a zoomed-in view of the ebractenoids clusters in Fig. 7(e). Figure 7(e) shows that noisy HSQC spectra are clustered close to their original spectrum, and thus, noise to the levels we have evaluated, has little effect on the ability of SMART to accurately place compounds into their appropriate location (ebractenoids in this case). Selected noise maps are provided in the Supplementary Information.

CONCLUSIONS AND FUTURE WORK

SMART is the first combination of NUS 2D NMR and deep CNNs. This tool streamlines dereplication and determination of natural product families from multiple organisms and facilitates their isolation and structural elucidation. While compound families represented the metadata associated with HSQC spectra in this study, it is very possible to associate and integrate biological, pharmacological and ecological data with SMART, and thereby create new tools for enhanced discovery and development of biological active NPs as well as other small molecules. Ultimately, this leads to an increased appreciation for the structural diversity and therapeutic potential of natural products.

By both quantitative and qualitative inspection of SMART's cluster space, the following properties were suggested by the results: 1) the distance between nearby nodes of a clustering map is a measure of the structural similarity between compounds that share molecular moieties (e.g., functional groups, carbon skeletons, etc.), 2) chimeric compounds with structural features comprised of two independent families of compounds reside near or in between the component clusters (for example, saponins are located near and between other glycosides and terpenoids, in Fig. S2), 3) this accuracy of placement of new compounds in SMART should be enhanced as the size of the training set grows, 4) as the size of the training set increases for a given compound class, the accuracy of placement of a new test compound in that family improves, 5) even in the presence of random spectral noise, spectra are strongly associated to their structural chemical analogues. Nevertheless, the

accuracy of recognition correlating to the signal-to-noise ratio of HSQC spectra remains to be determined, as does the impact of solvent effects on chemical shifts or extraneous peaks appearing in the spectrum from electronic sources or impurities. As more compounds are added to the training set, the SMART system will naturally improve in accuracy and robustness, thereby accelerating natural product structural elucidation and thus drug discovery.

SMART has an immediate value in NP drug discovery efforts by providing rapid and automatic compound dereplication and assignment to molecular structure families. With further refinement of the SMART workflow, such as training for spectra of the same compound with different S/N ratios, deeper understanding of other parameters that impact spectral recognition, combining with other fast NMR techniques, SMART has the potential to enhance NPR and enable new directions of experimentation at the chemistry-biology interface.

REFERENCES

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. 79, 629–661 (2016).

2. Kursar, T. A. et al. Securing economic benefits and promoting conservation through bioprospecting. Bioscience 56, 1005–1012, https://doi.org/10.1641/0006-3568 (2006).

3. Liu, W. T. et al. MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in Streptomyces roseosporus. J. Antibiot. 67, 99–104, https://doi.org/10.1038/ja.2013.99 (2014).

4. Medema, M. H. et al. Minimum Information about a biosynthetic gene cluster. Nat. Chem. Biol. 11, 625–631 (2015).

5. Walsh, C. T. A chemocentric view of the natural product inventory. Nat. Chem. Biol. 11, 620–624 (2015).

6. Molinski, T. F. NMR of natural products at the 'nanomole-scale'. Nat. Prod. Rep. 27, 321–329, https://doi.org/10.1039/b920545b (2010).

7. Breton, R. C. & Reynolds, W. F. Using NMR to identify and characterize natural products. Nat. Prod. Rep. 30, 501–524, https://doi. org/10.1039/c2np20104f (2013).

8. Mobli, M., Maciejewski, M. W., Schuyler, A. D., Stern, A. S. & Hoch, J. C. Sparse sampling methods in multidimensional NMR. Phys. Chem. Chem. Phys. 14, 10835–10843, https://doi.org/10.1039/c2cp40174f (2012).

9. Kazimierczuk, K. & Orekhov, V. Y. Accelerated NMR spectroscopy by using compressed sensing. Angewandte Chemie-International Edition 50, 5556–5559, https://doi.org/10.1002/anie.201100370 (2011).

10. Palmer, M. R. et al. Sensitivity of nonuniform sampling NMR. J. Phys. Chem. B 119, 6502–6515, https://doi.org/10.1021/jp5126415 (2015).

11. Hyberts, S. G., Arthanari, H. & Wagner, G. Applications of non-uniform sampling and processing. Top. Curr. Chem. 316, 125–148, https://doi.org/10.1007/128_2011_187 (2012).

12. Hyberts, S. G., Milbradt, A. G., Wagner, A. B., Arthanari, H. & Wagner, G. Application of iterative soft thresholding for fast reconstruction of NMR data non-

uniformly sampled with multidimensional Poisson Gap scheduling. J. Biomol. Nmr 52, 315–327, https://doi.org/10.1007/s10858-012-9611-z (2012).

13. Maciejewski, M. W., Mobli, M., Schuyler, A. D., Stern, A. S. & Hoch, J. C. Data sampling in multidimensional NMR: fundamentals and strategies. Top. Curr. Chem. 316, 49–77, https://doi.org/10.1007/128_2011_185 (2012).

14. Robinette, S. L. et al. Hierarchical alignment and full resolution pattern recognition of 2D NMR spectra: application to nematode chemical ecology. Anal. Chem. 83, 1649–1657, https://doi.org/10.1021/ac102724x (2011).

15. Smurnyy, Y. D., Blinov, K. A., Churanova, T. S., Elyashberg, M. E. & Williams, A. J. Toward more reliable C-13 and H-1 chemical shift prediction: A systematic comparison of neural-network and least-squares regression based approaches. J. Chem. Inf. Model 48, 128–134, https://doi.org/10.1021/ci700256n (2008).

16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444, https://doi.org/10.1038/nature14539 (2015).

17. Schmidhuber, J. Deep learning in neural networks: An overview. Neural Networks 61, 85–117, https://doi.org/10.1016/j. neunet.2014.09.003 (2015).

18. Gerwick, W. H. et al. Structure of Curacin-a, a novel antimitotic, antiproliferative, and brine shrimp toxic natural product from the marine cyanobacterium Lyngbya majuscula. J. Org. Chem. 59, 1243–1245, https://doi.org/10.1021/jo00085a006 (1994).

19. Yoo, H. D. & Gerwick, W. H. Curacins B and C, new antimitotic natural products from the marine cyanobacterium Lyngbya majuscula. J. Nat. Prod. 58, 1961–1965, https://doi.org/10.1021/np50126a029 (1995).

20. Marquez, B., Verdier-Pinard, P., Hamel, E. & Gerwick, W. H. Curacin D, an antimitotic agent from the marine cyanobacterium Lyngbya majuscula. Phytochemistry 49, 2387–2389 (1998).

21. Tarsis, E. M., Rastelli, E. J., Wengryniuk, S. E. & Coltart, D. M. The apratoxin marine natural products: isolation, structure determination, and asymmetric total synthesis. Tetrahedron 71, 5029–5044, https://doi.org/10.1016/j.tet.2015.05.047 (2015).

22. Choi, H., Mevers, E., Byrum, T., Valeriote, F. A. & Gerwick, W. H. Lyngbyabellins K-N from two Palmyra Atoll collections of the marine cyanobacterium Moorea bouillonii. Eur. J. Org. Chem., 5141–5150; https://doi.org/10.1002/ejoc.201200691 (2012).

23. Marner, F. J., Moore, R. E., Hirotsu, K. & Clardy, J. Majusculamides A and B, 2 epimeric lipodipeptides from Lyngbya majuscula Gomont. J. Org. Chem. 42, 2815–2819, https://doi.org/10.1021/jo00437a005 (1977).

24. Carter, D. C., Moore, R. E., Mynderse, J. S., Niemczura, W. P. & Todd, J. S. Structure of majusculamide-C, a cyclic depsipeptide from Lyngbya majuscula. J. Org. Chem. 49, 236–241, https://doi.org/10.1021/jo00176a004 (1984).

25. Moore, R. E. & Entzeroth, M. Majusculamide-D and deoxymajusculamide-D, two cytotoxins from Lyngbya majuscula. Phytochemistry 27, 3101–3103, https://doi.org/10.1016/0031-9422(88)80008-6 (1988).

26. Bodis, L., Ross, A., Bodis, J. & Pretsch, E. Automatic compatibility tests of HSQC NMR spectra with proposed structures of chemical compounds. Talanta 79, 1379–1386, https://doi.org/10.1016/j.talanta.2009.06.017 (2009).

27. Hinneburg, A., Egert, B. & Porzel, A. Duplicate detection of 2D-NMR Spectra. Journal of Integrative Bioinformatics 4, 64, https://doi. org/10.2390/biecoll-jib-2007-53 (2007).

28. Wolfram, K., Porzel, A. & Hinneburg, A. Similarity search for multi-dimensional NMR-spectra of natural products. Knowledge Discovery in Databases: Pkdd 2006, Proceedings 4213, 650–658 (2006).

29. Levitt, M. H. Spin Dynamics: Basics of Nuclear Magnetic Resonance, 2nd edn, 345 (John Wiley & Sons, 2008).

30. Chopra, S., Hadsell, R. & LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. Proc. CVPR. IEEE., 539–546 (2005).