Using Deep Siamese Neural Networks to Speed up Natural Products Research

Nicholas Roberts[1]
nick11roberts@cmu.edu

Poornav S. Purushothama[7]
poornavsargoor@gmail.com

Vishal T. Vasudevan[3]
vasuvish@amazon.com

Siddarth Ravichandran[2]
s2ravich@eng.ucsd.edu

Chen Zhang[2, 4, 5]
beowulf.zc@gmail.com

William H. Gerwick[4, 5, 6]
wgerwick@ucsd.edu

Garrison W. Cottrell[2]
gary@ucsd.edu

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Computer Science and Engineering, University of California, San Diego
[3]Amazon
[4]Center for Marine Biotechnology and Biomedicine, University of California, San Diego
[5]Scripps Institution of Oceanography, University of California, San Diego
[6]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
[7]Hewlett Packard Enterprise Company

Abstract

Natural products (NPs) are an important source of novel disease treatments. A bottleneck in the search for new NPs is structure determination of molecules extracted from biological organisms. One method is to use 2D Nuclear Magnetic Resonance (NMR) spectroscopy, which indicates bonds between nuclei in the compound and hence is the "fingerprint" of the compound. Computing a similarity score between 2D NMR spectra for a novel compound and a compound whose structure is known provides clues to the structure of the novel compound. Standard approaches to this problem do not scale to larger databases of compounds. Here we use deep convolutional Siamese networks to map NMR spectra to a cluster space, where similarity is given by the distance in the space. This approach results in an AUC score that is more than four times better than an approach using LDA.

1 Introduction

Natural products (NPs) obtained from both terrestrial and marine organisms are the single most important source of drug leads and new therapeutics. Approximately 70% of all drugs in the clinic today have an origin or inspiration from natural products of plants, animals and microorganisms (Gerwick & Moore; Mayer et al.; Molinski et al.; Newman & Cragg). NPs have also been a major inspiration for the development of many of the pharmaceutical drugs currently available. This trend is continuing in that NPs, NP derivatives and their mimics, account for roughly 50% of all new drugs over the past several years (Newman & Cragg). Thus, NPs continue to be an important source of new pharmaceuticals and pharmaceutical leads (Pye et al.).

A major bottleneck in drug discovery is determining the molecular structure of a novel substance. This process is quite time consuming, and although a skilled and experienced NP researcher can be quite effective in this pursuit, most structure elucidations are limited by the poor quality of data,

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

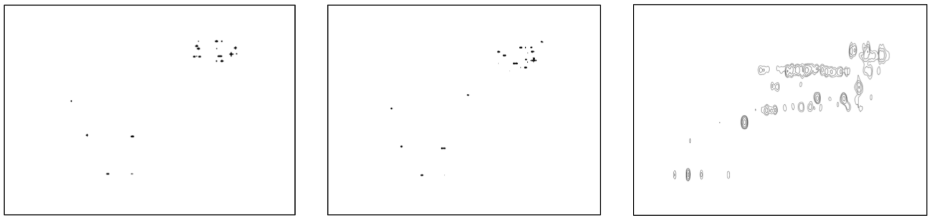

(a) Aspewentin A (b) Aspewentin B (c) Chandonanone A

Figure 1: Examples of 2D NMR spectra, molecular “fingerprints.”

subjective human evaluation of these data, and a step-wise deduction of the structure indicated by this information. Therefore, the structure elucidation of novel chemical entities is oftentimes a long and laborious process. To effectively mine the rich repertoire of chemical diversity of NPs, as well as to identify the diverse NPs present in various other classes of organisms, new automated methods (Leao et al.; Wang et al.; Zhang et al., b) of structural analysis are greatly needed. In fact, some cheminformatics tools developed to meet this need have achieved wide application (Taboada et al.; Tao et al.; Zhang et al., a).

A major technique used for this process is 2D Nuclear Magnetic Resonance (NMR) spectroscopy, and in particular, Heteronuclear single-quantum correlation spectroscopy (HSQC NMR) has been found to be work well for small molecules. This method detects correlations between nuclei of two different types that are linked by one bond. This gives one peak per pair of coupled nuclei, whose two coordinates are the chemical shifts of the two coupled atoms. Three examples are shown in Figure 1. The first two are from the same compound family, and so their spectra are similar. This approach requires very careful analysis that is time consuming, and requires a high level of investigator skill and experience. Our goal is to use machine learning to help accelerate this process. In short, the idea is to learn to map novel NMR spectra into a similarity space where compounds with known structures similar to the novel compound are nearby in the space, thus giving cues to its structure. This technique can also tell us whether the candidate compound is actually one we have seen before, a process that is awkwardly named “dereplication” in this field.

In this paper, we apply deep siamese convolutional neural networks to this problem and demonstrate improved performance over machine learning techniques previously applied to this problem, specifically, Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA) (Hofmann; Blei et al.).

2 Related work

A number of techniques are currently being employed for the purpose of performing similarity search of one-dimensional NMR spectra (Barros & Rutledge, 2005; Krishnan et al., 2004; Steinbeck et al., 2003). However, there are very few similarity search techniques for two-dimensional NMR spectra. The most significant technique was proposed by Wolfram et al. (2006). Their approach models two dimensional NMR spectra as documents, and uses topic modeling methods to find similarity between NMR spectra. Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) is used to find the hidden topics in the documents.

3 Methods

3.1 Data acquisition and processing

We collected 4,105 HSQC NMR spectra from the *Journal of Natural Products*, and *Organic Letters*, years 2011 through 2015. The distribution of compounds is heavily skewed toward smaller families, with many families having fewer than 10 examples. We found that using families with fewer than 10 examples resulted in too much noise, so we removed these samples. The remaining dataset contained 1,385 NMR spectra from 104 families. We refer to this subset of the data as SMART10, and all of our experiments were done using this dataset.

3.2 Methods of comparison

To evaluate how well our Siamese CNN performed, we used two topic modeling techniques: PLSI, as used by Wolfram et al., and Latent Dirichlet Allocation (LDA) (Blei et al.) to measure baseline performance on our dataset. We first mapped our NMR spectra to documents and words as described in Wolfram et al. (2006) to use the topic modeling techniques.

3.3 Architecture

For this task we use a deep siamese convolutional neural network to learn a metric space of raw NMR spectra interpreted as images (Bromley et al., 1993; Chopra et al., 2005; Hadsell et al., 2006). When all of the training data is projected into this metric space, the result is a searchable cluster map of NMR spectra.

Siamese neural networks, first introduced by Bromley et al. (1993), have been shown to be successful in situations where the number of classes is not known at the time of training, when the number of classes is very large, and when the number of examples per class is very small.

Here, we counted any substances from the same family as "the same" and any from different families as different. The network uses an objective function that reduces the distance between same examples and expands it between different ones. Thus the network clusters the inputs in the output space - there is no specific target that the network is trying to achieve. The dimensionality of the output of the network determines the dimensionality of the cluster space. For example, the Aspewentins (as in Figure 1) form a single family, so the network tries to map these NMR spectra to nearby points in the output space, and tries to map other compounds farther away. A variant of contrastive loss is used for our system, where a "pushing factor" P is introduced as a multiplicative scaling factor of the negative examples (we find that $P = 1.5$ works best).

$$L(W, Y, X_1, X_2) = (1 - Y)\frac{1}{2}(D_W(X_1, X_2))^2 + (Y)\frac{P}{2}\{max(0, m - D_W(X_1, X_2))\}^2, \tag{1}$$

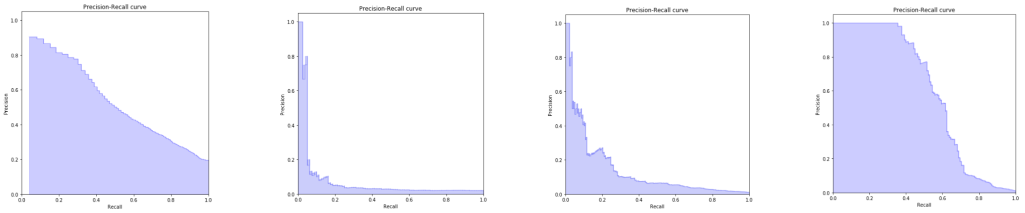
Our architecture consists of four convolutional layers with unit stride and maxpooling ($4 \times 4 \times 8$, $7 \times 7 \times 16$, $4 \times 4 \times 16$, $4 \times 4 \times 16$), and four fully connected layers with dropout (0.5) of dimensions 128, 128, 128, and k where K is the dimension of the similarity metric space. This architecture is replicated with tied weights so as to be used in the framework of siamese networks.

3.4 Experiments

We performed two experiments. In the first experiment, we separated the training, holdout, and test sets by randomly splitting the 1,385 NMR spectra into 80% training, 10% validation, and 10% test sets. In this case, the compounds in the holdout and test sets will generally have compounds in the same family in the training set. We use the loss on the validation set to stop training, and then report the precision recall curves and the AUC for the test set. All three approaches were evaluated this way.

The second experiment addresses the actual use-case of the model, where a completely novel set of compounds is tested. To model this case, we held out four families from training for use in testing: the aphanamixoids, teuvissides, tasiamides, and macrolactins. We call these the "probe families." Evaluation here is more difficult, as there is no ground truth for the result.

Instead, in this case we calculate the averaged Tanimoto score (a measure of structure similarity, closely related to the Jaccard Similarity coefficient, scaled between 0 and 100) for the top five closest compound families of the probe families, using the PubChem Score Matrix Service (Cai et al.; Lv et al.; Mevers et al.; Mondol et al.; Kim et al.). Ideally, the PubChem website assigns each chemical compound a unique PubChem Compound Identification (CID) number in order to run the Tanimoto score calculation. Unfortunately, not all CIDs of the compounds are retrievable. Furthermore, we observed compounds with incorrect structures in their CID database (e.g., munronins) (Yan et al., 2015). Hence, we only performed the Tanimoto score calculation for aphanamixoids and teuvissides, comparing them to some of their top hits provided by our system whose CIDs were accessible on PubChem.



(a) PLSI (small dataset) (b) PLSI (NAT10) (c) LDA (NAT10) (d) Siamese (NAT10)

Figure 2: Precision-recall curves for PLSI, LDA, and the Siamese network. Panel (a) is a demonstration that our implementation of PLSI obtains results similar to those in their paper with a similar-sized dataset.

4 Results

4.1 Performance on held out data

In order to verify our implementation of the grid-based vocabulary and PLSI, we aimed to replicate the results of Wolfram et al. using the same amount of data and the same cross-validation technique as used in their work (Wolfram et al., 2006). However, using the full SMART10 dataset, we find that PLSI performs quite poorly compared to scenarios in which there are fewer classes and significantly less data. We find that LDA outperforms PLSI on the SMART10 dataset. LDA has higher precision at most recall values, whereas PLSI only has high precision when recall is low. We find that LDA has an AUC score of 0.14, which is a slight improvement over the AUC score achieved by PLSI.

Our siamese architecture dramatically outperforms LDA and PSLI There is a significant difference in the performance displayed by these curves in favor of our method. Concretely, we find that the AUC score for PLSI is 0.09, while the AUC score for our model is 0.60. The advantage of the learned representation using the Siamese loss is obvious.

4.2 Performance on novel compound families

As mentioned above, one particular advantage of siamese neural networks is the ability to generalize to new classes that it has not been trained on, and in fact, the number of classes that it is expected to support does not have to be known at training time. This property is of considerable interest to us because the real-world domain is such that the number of all classes, compound families in this case, is unknown.

Table 1 shows the top five hits for two compound families, and the similarity (inverse distance in the cluster space) between them and the held out family. Tanimoto scores of 2D chemical structures were calculated in comparison with the similarity score of the NMR spectra of those compounds generated by the siamese neural network model (see Table 3 for the top 5 hits list).

A higher Tanimoto score indicates structure similarity of two compounds. The average intra-cluster Tanimoto score of the cluster containing aphanamixoids (aph.) C, D, E, F and G is 95.7, and the cluster containing turrapubins (turr.) A, B, C, D, E, F, G, H, I and J is 84.5. The average intra-cluster Tanimoto score of the cluster containing khayseneganins (khay.) D, H and 3-deacetylkhivorin is 90.9 (Yuan et al.). All of these intra-cluster Tanimoto scores are higher than the inter-cluster Tanimoto score $T_{(aph.-turr.)} = 70$ or $T_{(aph.-khay.)} = 74$. However, in contrast to the Tanimoto scores, turrapubins are ranked closer to aphanamixoids than khayseneganins by our system, which is consistent with the fact that β-furan linked lactone is present in khayseneganins, whereas β-furan is directly linked to cyclopentane derivatives in the other two cases. These results indicate that our system is capable of associating unknown compounds to their known analogues in our training dataset, consistent with human judgment.

By the same token, the average intra-cluster Tanimoto score of the cluster containing teuvissides (teuv) A, B, C, D, E, F, G and H is 99.3, and the cluster containing oleraceins (oleo) O, K and L is 99.6 (Lv et al.; Jiao et al., 2015). The average intra-cluster Tanimoto score of the cluster containing sophodibenzosides (soph) A and B is 98.0, and the cluster containing flemingins (flem) A, B, C and O is 98.5 (Shen et al., 2013; Gumula et al., 2014). In contrast, the inter-cluster Tanimoto scores of the four clusters are $T_{(teuv.-oler.)} = 68$, $T_{(teu.-sophod.)} = 65$ or $T_{(teuv.-flem.)} = 57$. In this case, the

Table 1: Top 5 hits for 2 held out families.

HELD OUT FAMILY	APHANAMIXOIDS	SIMILARITY	TEUVISSIDES	SIMILARITY
Rank 1	turrapubins	3.56	oleraceins	1.30
Rank 2	munronins	1.52	sophodibenzoside s	1.28
Rank 3	khayseneganins	1.42	aquaterins	1.15
Rank 4	Inositol Derivatives	1.23	bruceollines and yadanziolides	1.09
Rank 5	pedinophyllols	1.20	flemingins	1.01

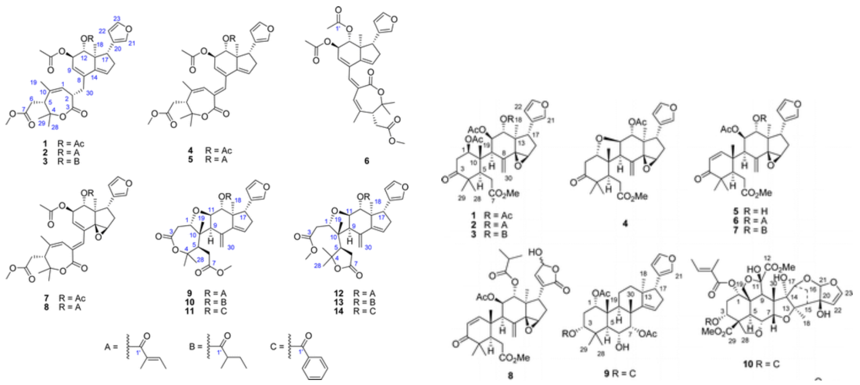


Figure 3: Chemical structures for Aphanamixoids and Turrapubins, respectively

Tanimoto similarity score matches the results of our system. Empirically, teuvissides and oleraceins are both glycosylated coumaroyltyramines, while sophodienzosides are dibenzoyl glycosides. Again, our system works nicely with respect to detecting glycosides.

Therefore, our system not only can cluster HSQC NMR spectra based entirely on the chemical structure similarity, but also outperforms the Tanimoto algorithm in structure similarity scoring.

5 Conclusions and future work

We have shown that siamese neural networks perform significantly better than PLSI and LDA for the task of learning a similarity metric of NMR spectra to assist structure determination in Natural Product research. Our model maps directly from NMR spectra into a cluster space, where nearby points in the space have similar chemical structures. This architecture allows newly-discovered compounds to be mapped into the same space, resulting in a list of similarly-structured compounds. This list of similar compounds then provides strong clues to the structure of the novel compound, reducing the number of further experiments required to identify the structure.

In future work, we will apply hyperparameter optimization techniques to improve the architecture. In particular, we will apply the Bayesian hyperparameter tuning system Hyperopt[1]. We will also explore using an autoencoder on top of the cluster space, which is known to improve performance (Yann LeCun, personal communication).

References

António S Barros and Douglas N Rutledge. Segmented principal component transform–principal component analysis. *Chemometrics and intelligent laboratory systems*, 78(1-2):125–137, 2005.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International*

[1] https://github.com/hyperopt/hyperopt

Conference on Neural Information Processing Systems, NIPS'93, pp. 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=2987189.2987282.

J. Y. Cai, D. Z. Chen, S. H. Luo, N. C. Kong, Y. Zhang, Y. T. Di, Q. Zhang, J. Hua, S. X. Jing, S. L. Li, S. H. Li, X. J. Hao, and H. P. He. Limonoids from aphanamixis polystachya and their antifeedant activity. *Journal of Natural Products*, 77(3):472–482.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pp. 539–546, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.202. URL http://dx.doi.org/10.1109/CVPR.2005.202.

W. H. Gerwick and B. S. Moore. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology (vol 19, pg 85, 2012). *Chemistry & Biology*, 19(12): 1631–1631.

I. Gumula, J. P. Alao, I. O. Ndiege, P. Sunnerhagen, A. Yenesew, and M. Erdelyi. Flemingins g-o, cytotoxic and antioxidant constituents of the leaves of flemingia grahamiana. *J Nat Prod*, 77(9): 2060–7, 2014. ISSN 1520-6025 (Electronic) 0163-3864 (Linking). doi: 10.1021/np500418n. URL http://www.ncbi.nlm.nih.gov/pubmed/25226568. Gumula, Ivan Alao, John Patrick Ndiege, Isaiah Omolo Sunnerhagen, Per Yenesew, Abiy Erdelyi, Mate eng Research Support, Non-U.S. Gov't 2014/09/17 06:00 J Nat Prod. 2014 Sep 26;77(9):2060-7. doi: 10.1021/np500418n. Epub 2014 Sep 16.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pp. 1735–1742, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.100. URL http://dx.doi.org/10.1109/CVPR.2006.100.

T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42 (1-2):177–196.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 50–57, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312649. URL http://doi.acm.org/10.1145/312624.312649.

Z. Z. Jiao, S. Yue, H. X. Sun, T. Y. Jin, H. N. Wang, R. X. Zhu, and L. Xiang. Indoline amide glucosides from portulaca oleracea: Isolation, structure, and dpph radical scavenging activity. *J Nat Prod*, 78(11):2588–97, 2015. ISSN 1520-6025 (Electronic) 0163-3864 (Linking). doi: 10.1021/acs.jnatprod.5b00524. URL http://www.ncbi.nlm.nih.gov/pubmed/26562741. Jiao, Ze-Zhao Yue, Su Sun, Hong-Xiang Jin, Tian-Yun Wang, Hai-Na Zhu, Rong-Xiu Xiang, Lan eng Research Support, Non-U.S. Gov't 2015/11/13 06:00 J Nat Prod. 2015 Nov 25;78(11):2588-97. doi: 10.1021/acs.jnatprod.5b00524. Epub 2015 Nov 12.

S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Y. Han, J. E. He, S. Q. He, B. A. Shoemaker, J. Y. Wang, B. Yu, J. Zhang, and S. H. Bryant. Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213.

P Krishnan, NJ Kruger, and RG Ratcliffe. Metabolite fingerprinting and profiling in plants using nmr. *Journal of experimental botany*, 56(410):255–265, 2004.

T. Leao, G. Castelao, A. Korobeynikov, E. A. Monroe, S. Podell, E. Glukhov, E. E. Allena, W. H. Gerwick, and L. Gerwick. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus moorea. *Proceedings of the National Academy of Sciences of the United States of America*, 114(12):3198–3203.

H. W. Lv, M. D. Zhu, J. G. Luo, and L. Y. Kong. Antihyperglycemic glucosylated coumaroyltyramine derivatives from teucrium viscidum. *Journal of Natural Products*, 77(2):200–205.

A. M. S. Mayer, K. B. Glaser, C. Cuevas, R. S. Jacobs, W. Kem, R. D. Little, J. M. McIntosh, D. J. Newman, B. C. Potts, and D. E. Shuster. The odyssey of marine pharmaceuticals: a current pipeline perspective. *Trends in Pharmacological Sciences*, 31(6):255–265.

E. Mevers, F. P. J. Haeckl, P. D. Boudreau, T. Byrum, P. C. Dorrestein, F. A. Valeriote, and W. H. Gerwick. Lipopeptides from the tropical marine cyanobacterium symploca sp. *Journal of Natural Products*, 77(4):969–975.

T. F. Molinski, D. S. Dalisay, S. L. Lievens, and J. P. Saludes. Drug development from marine natural products. *Nature Reviews Drug Discovery*, 8(1):69–85.

M. A. M. Mondol, F. S. Tareq, J. H. Kim, M. A. Lee, H. S. Lee, Y. J. Lee, J. S. Lee, and H. J. Shin. Cyclic ether-containing macrolactins, antimicrobial 24-membered isomeric macrolactones from a marine bacillus sp. *Journal of Natural Products*, 74(12):2582–2587.

D. J. Newman and G. M. Cragg. Natural products as sources of new drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661.

C. R. Pye, M. J. Bertin, R. S. Lokey, W. H. Gerwick, and R. G. Linington. Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences of the United States of America*, 114(22):5601–5606.

Y. Shen, Z. M. Feng, J. S. Jiang, Y. N. Yang, and P. C. Zhang. Dibenzoyl and isoflavonoid glycosides from sophora flavescens: inhibition of the cytotoxic effect of d-galactosamine on human hepatocyte hl-7702. *J Nat Prod*, 76(12):2337–45, 2013. ISSN 1520-6025 (Electronic) 0163-3864 (Linking). doi: 10.1021/np400784v. URL `http://www.ncbi.nlm.nih.gov/pubmed/24295087`. Shen, Yi Feng, Zi-Ming Jiang, Jian-Shuang Yang, Ya-Nan Zhang, Pei-Cheng eng Research Support, Non-U.S. Gov't 2013/12/04 06:00 J Nat Prod. 2013 Dec 27;76(12):2337-45. doi: 10.1021/np400784v. Epub 2013 Dec 2.

Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of chemical information and computer sciences*, 43(6):1733–1739, 2003.

C. Taboada, A. E. Brunetti, F. N. Pedron, F. C. Neto, D. A. Estrin, S. E. Bari, L. B. Chemes, N. P. Lopes, M. G. Lagorio, and J. Faivovich. Naturally occurring fluorescence in frogs. *Proceedings of the National Academy of Sciences of the United States of America*, 114(14):3672–3677.

Y. W. Tao, P. L. Li, D. J. Zhang, E. Glukhov, L. Gerwick, C. Zhang, T. F. Murray, and W. H. Gerwick. Samholides, swinholide-related metabolites from a marine cyanobacterium cf. phormidium sp. *Journal of Organic Chemistry*, 83(6):3034–3046.

M. X. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Criisemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderon, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C. C. Liaw, Y. L. Yang, H. U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Q. Dai, R. Neupane, J. Gurr, A. M. C. Rodriguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P. M. Allard, P. Phapale, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837.

Karina Wolfram, Andrea Porzel, and Alexander Hinneburg. Similarity search for multi-dimensional nmr-spectra of natural products. In *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pp. 650–658, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-45374-1, 978-3-540-45374-1. doi: 10.1007/11871637_67. URL `http://dx.doi.org/10.1007/11871637_67`.

Ying Yan, Jian-Xin Zhang, Tao Huang, Xin-Ying Mao, Wei Gu, Hong-Ping He, Ying-Tong Di, Shun-Lin Li, Duo-Zhi Chen, Yu Zhang, et al. Bioactive limonoid constituents of munronia henryi. *Journal of natural products*, 78(4):811–821, 2015.

C. M. Yuan, G. H. Tang, Y. Zhang, X. Y. Wang, M. M. Cao, F. Guo, Y. Li, Y. T. Di, S. L. Li, H. M. Hua, H. P. He, and X. J. Hao. Bioactive limonoid and triterpenoid constituents of turraea pubescens. *Journal of Natural Products*, 76(6):1166–1174.

C. Zhang, Y. Idelbayev, N. Roberts, Y. W. Tao, Y. Nannapaneni, B. M. Duggan, J. Min, E. C. Lin, E. C. Gerwick, G. W. Cottrell, and W. H. Gerwick. Small Molecule Accurate Recognition Technology (SMART) to enhance natural products research. *Scientific Reports*, 7, a.

C. Zhang, C. B. Naman, N. Engene, and W. H. Gerwick. Laucysteinamide a, a hybrid pks/nrps metabolite from a saipan cyanobacterium, cf. caldora penicillata. *Marine Drugs*, 15(4), b.