Cite this: *Med. Chem. Commun.*,
2019, 10, 1667

Received 2nd March 2019,
Accepted 4th June 2019

DOI: 10.1039/c9md00128j

rsc.li/medchemcomm

Is it time for artificial intelligence to predict the function of natural products based on 2D-structure

Miaomiao Liu, ^a Peter Karuso, ^b Yunjiang Feng, ^a Esther Kellenberger, ^c
Fei Liu, ^b Can Wang ^d and Ronald J. Quinn ^{*a}

Currently, there is no established technique that allows the function of a compound produced by nature to be predicted by looking at its 2-dimensional chemical structure. One of chemistry's grand challenges: to find a function for every known metabolite. We explore the opportunity for Artificial Intelligence to provide rationale interrogation of metabolites to predict their function.

Introduction

The profound and specific biological activity of natural products coupled with their immediately recognizable 2D-structures suggests a code within these structures that we are not as yet aware of. The long-standing challenge is to be able to decode the functional information entangled in the 2D-structures of these metabolites, selected over millions of

years by continuous evolution. The function of natural products could be biological, ecological, pharmacological or to influence metabolite production. Almost all of these functions are the result of an interaction with a macromolecule, predominantly proteins.

An indication of the challenge to correlate the function of a natural product with its 2D-structure is provided by natural products isolated after screening against molecular targets. Some examples are given in Fig. 1. Sideroxylonal C (1) from *Eucalyptus albens* Benth. is an inhibitor of human plasminogen activator type-1 (PAI-1) and resulted from the screening of 21 384 extracts.¹ Adociasulfate 1 (2) inhibited the osteoclast vacuolar H⁺-ATPase proton pump in hen bone-derived membrane vesicles.² 25-Hydroxy-13(24),15,17-cheilanthatrien-19,25-olide (3) was one of four cheilanthane sesterterpenes to inhibit mitogen and stress activated kinase (MSK1).³ Dysinosin

^a Griffith Institute for Drug Discovery, Griffith University, Brisbane, Qld 4111, Australia. E-mail: r.quinn@griffith.edu.au; Tel: +61 7 3735 6006

^b Department of Molecular Sciences, Macquarie University, Sydney, NSW, 2109, Australia

^c Laboratory of Therapeutic Innovation, Medalis Drug Discovery Center, University of Strasbourg, Illkirch, France

^d School of Information and Communication Technology, Griffith University, Gold Coast campus, Qld 4222, Australia

Yunjiang Feng and Miaomiao Liu

for Drug Discovery. Her research interests include bioactive natural products and traditional Chinese medicine.

Miaomiao Liu (right) obtained dual PhD degrees in 2017 from Griffith University supervised by Ronald J Quinn and the University of Chinese Academic of Sciences supervised by Lixin Zhang. She is now a Research Fellow at the Griffith Institute for Drug Discovery, Griffith University. Her research interests involve the target identification using native mass spectrometry, identification of bioactive natural products using a pheno-target approach and NMR fingerprints.

Yunjiang Feng (left) obtained a Bachelor's degree in Pharmaceutical Chemistry from Peking University (former Beijing Medical University), then a PhD degree in marine natural products chemistry from James Cook University, followed by post-doctoral research in bioactive fungal metabolites in Canterbury University. In 2004, Dr Feng was recruited as a research fellow to Griffith University. She is currently an Associate Professor. Dr Feng current leads a research group at Griffith Institute

A (4) from a marine sponge of the family Dysideidae was found to be a potent inhibitor of the blood coagulation cascade factor VIIa.^{4,5} Forty thousand (40 000) extracts were screened against *Helicobacter pylori* aspartate semialdehyde dehydrogenase (ASD) resulting in the identification of petrosamine B (5), as an inhibitor of the enzyme.⁶ Latifolians A (6) and B were new examples of the 8-benzyl-berberine alkaloid structure class and resulted from the screening of approximately 100 000 extracts against the neuronal specific isoform of the c-Jun N-terminal kinases, JNK3. Both compounds inhibited the kinase.⁷ Grandisine A (7) is a novel indolizidine alkaloids with human δ -opioid receptor binding affinity.⁸ Endiandrin A (8) was found to be a potent glucocorticoid receptor (GR) binder.⁹ Stylissadines A (9) and B, were identified as specific antagonists of the ligand gated cation channel P2X₇ receptor.¹⁰ The stylissadines were isolated from the Aus-

tralian marine sponge *Stylissa flabellata* Ridley & Dendy 1886 and are bisimidazo-pyrano-imidazole bromopyrrole ether alkaloids. Determination of the absolute configuration suggested that a number of related natural products, including palau'amine, should be revised to 12*R*, 17*S*, 20*S*. Isoprenylcysteine carboxyl methyltransferase (Icmt) catalyses the carboxyl methylation of oncogenic proteins and an extract from a *Pseudocerotina* sp. was identified in a HTS campaign. Spermatamine (10) is the first natural product inhibitor of isoprenylcysteine carboxyl methyltransferase.¹¹ Lysianadioic acid (11), is a potent inhibitor of carboxypeptidase B (CPB) and is a new arginine analogue containing an unusual dicarboxylic acid.¹² Exiguaquinol (12) is a novel pentacyclic hydroquinone from *Neopetrosia exigua* that inhibits *Helicobacter pylori* Murl, a glutamate reaceamase.¹³ Clavatadine A (13), a natural product with selective recognition and irreversible

Peter Karuso and Fei Liu

Peter Karuso obtained his BSc(Hon 1) and PhD from the Department of Organic Chemistry at the University of Sydney (W. C. Taylor) and completed postdoctoral fellowships with Dame Patricia Bergquist (University of Auckland), Paul Scheuer (University of Hawaii), Ian Scott and Sir Derek Barton (Texas A&M) and Horst Kessler (Technische Universität München) before returning to Macquarie University in 1990 and is currently Professor of Chemistry. His research interests are in understanding the role and function of natural products in biological systems using "reverse chemical proteomics" to rapidly, and agnostically, link natural products with their protein targets using phage and yeast surface display technologies.

Fei Liu did her undergraduate studies in Chemistry at John Carroll University and her PhD in Organic Chemistry at Yale University. After an NIH postdoctoral fellowship on "unnatural" products from biosyn-

thetic pathways at the Harvard Medical School, she moved from Boston to Sydney in 2004 and is currently Senior Lecturer in the Department of Molecular Sciences at Macquarie University. Her long-term interest in natural product-driven chemical proteomics for basic biological research and drug discovery is being pursued in close collaboration with the Australian Proteome Analysis Facility (APAF) and reverse chemical proteomics discovery platforms at Macquarie.

Esther Kellenberger

Esther Kellenberger obtained a Ph.D. (with Bruno Kieffer) in biophysics from Strasbourg University, France, in 2000. She was a Von Humboldt fellow (with Michael Sattler) at the European Molecular Biology Laboratory, Germany, in 2001. She has since worked in the faculty of Pharmacy, Strasbourg University, where she is Professor of computer-aided drug design. Her research activities are focused on the structure-based discovery

of active molecules. She is particularly interested in understanding and modeling molecular recognition (mining of the Protein Data Bank, docking and site comparison methods for virtual screening, function of the chemokine receptor CCR5).

Can Wang

Can Wang received her B.Sc. and M.Sc. degrees in Mathematics from Wuhan University, China, in 2007 and 2009, respectively, and the Ph.D. degree on Computing Sciences from the Advanced Analytics Institute, University of Technology, Sydney, NSW, Australia, in 2013. She worked as a postdoctoral fellow with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia in Hobart, Tasmania from 2014 to 2016.

She is currently a lecturer in the Gold Coast campus, Griffith University, Australia. Her current research interests include data analytics, artificial intelligence, machine learning, and big data.

inhibition of factor XIa was isolated from a marine sponge, *Suberea clavata* Pulitzer-Final 1982.¹⁴ The first example of screening extracts using native mass spectrometry in an electrospray ionization Fourier transform ion cyclotron resonance mass spectrometer identified 6-(1S-hydroxy-3-methyl-butyl)-7-methoxy-2H-chromen-2-one (14) as the bovine carbonic anhydrase II active compound.¹⁵ The resveratrol tetramer (-)-hopeaphenol (15) inhibits type III secretion in the gram-negative pathogens *Yersinia pseudotuberculosis* and *Pseudomonas aeruginosa*.¹⁶ Euodenine A (16) is a small molecule agonist of human Toll-Like receptor 4 (TLR4) isolated from the leaves of *Euodia asteridula*.¹⁷ Euodenine A is a human-selective agonist that is CD14-independent and requires both TLR4 and MD-2 for full efficacy, and could modulate the Th2 immune response without causing lung damage. Venuloside A (17) from *Pittosporum venulosum* targets the LAT3 amino acid transporter.¹⁸ Achyrodimier F (18) is a tyrosyl-DNA phosphodiesterase I (Tdp1) inhibitor.¹⁹

Target based screening results in the identification of many bioactive natural products such as those discussed above (Fig. 1). However, the investment is large, the process is inefficient and results only in ligands for known targets with little or no ability to predict the function of any other natural product.

The problem is even more intense if the natural product is isolated against a cellular target. Some examples are shown in Fig. 2. In this case, target identification is a major difficulty. 1-Methyl isoguanosine (19) was isolated from the aqueous ethanolic extract of the marine sponge *Tedania digitata* and was later shown to be a non-selective agonist at Adenosine A1 and A2A receptors.^{20–24} Axinellamines A (20) and Axinellamines B-D are imidazo-azole-imidazole bromopyrroloes isolated from the Australian marine sponge, *Axinella* sp. They had weak bactericidal activity against the screening organism *Helicobacter pylori*.²⁵ The axinellamines were later synthesised by Baran *et al.* in a scalable process to

allow wider testing and found to have significant anti-bacterial activity including against both hospital-acquired and community-acquired methicillin-resistant *Staphylococcus aureus* (MRSA) and Gram-negative bacteria. Iotrochotazine A (21) had cellular effects on EEA1-associated early endosomes together with decreased lysosomal staining on human olfactory neurosphere derived cells (hONS) from Parkinson's disease patients.²⁶ Jaspamycin (22) had the highest deviation from control over the 38 biological parameters in the unbiased hONS cell phenotypic assay out of the 22 secondary metabolites isolated from *Jaspis splendens*.²⁷ Subsequent target identification is required for molecules that have cellular activity.

Fragment-based screening using low molecular weight (MW) natural products gives information on local binding sites within proteins. Our fragment-based publication identified 96 natural product fragments as binding partners of 32 of the putative malarial targets (Fig. 3).²⁸

We review data that may be analysed by Artificial Intelligence (AI) to answer the grand challenge: to find a function for every known metabolite by providing annotation of natural product interaction with proteins.

Discussion

Six datasets offering a combination of information related to known correlations of natural products to molecular target, structure similarity/dissimilarity, scaffolds and phenotypic cellular activity could be used in AI algorithms (Fig. 4). The objective is to annotate every natural product with its molecular target(s). The six datasets are now discussed individually.

Dataset 1. Molecular target

Molecular targets have been discovered for a number of natural products and a comprehensive literature search can be used to populate this dataset. All known targets would be included against each compound. By way of example this dataset would contain the PDB code 3BG8 for clavadin A (in complex with Factor XIa).¹⁴ Native mass spectrometry can be used to identify protein-ligand complexes. The technology is robust and relies on non-denaturing electrospray-ionization (ESI) to firstly recognize multi-charged proteins in their near-native states. High resolution, high mass accuracy measurements, coupled with soft ionization techniques to preserve the integrity of complexes, allows for the confirmation of protein targets of natural products (Fig. 4A).^{15,29}

Dataset 2. Molecular targets from fragment-based screening

Fragment-based screening uses low MW compounds to identify multiple binding sites within a protein. There are over 20 000 natural products that have MW < 250. We used native mass spectrometry to investigate 62 *Plasmodium falciparum* proteins as potential targets for antimalarial drugs using a natural product-based fragment library. We discovered 96 low molecular natural products identified as binding

Ron Quinn

Ron Quinn received his B.Sc. (Hons 1) and Ph.D. (with Ken Cavill) from UNSW. After post-doctoral training with Bob Pettit at Arizona State University, Richard Moore and Ted Norton at the University of Hawaii, and Rod Rickards at ANU he joined the Roche Research Institute for Marine Pharmacology in Sydney in 1974. He commenced at Griffith University in 1982. He is currently Professor of Chemistry.

His interests focus on bio-discovery for tuberculosis, malaria, and Parkinson's disease; fragment-based drug discovery (using low MW natural products); native state mass spectrometry; NMR metabolomics; medicinal chemistry and artificial intelligence.

Fig. 1 Natural products that have actions at protein targets. This illustrates the difficulty of predicting function from the chemical structure.

partners of 32 of the putative malarial targets. Seventy-nine (79) fragments had direct growth inhibition on *Plasmodium falciparum*.²⁸ There were 48 selective fragments that bound a single protein and 48 fragments that bound more than one protein.^{28,30}

Fig. 4B shows a typical network, visualized using Cytoscape software.³¹ Compounds are shown as space-filled representations, rectangular nodes represent proteins. Edge represents interaction between a fragment hit and a protein.

Considering hits and proteins as nodes, they are connected to each other based on the binding interaction, to produce a network.

Fragment-based networks gives 4 sets of data (Fig. 5). A fragment binding to a single protein provides a molecular target. A fragment that bound to 2 or more proteins would identify proteins that have common fragment-protein interactions. The native mass spectrometry experiment can be used for competitive binding and identify fragments that can

Fig. 2 Natural products that have actions at cellular targets. This illustrates the difficulty of predicting function from the chemical structure.

Fig. 3 Sixteen of 96 low MW natural products that bind to the malaria proteome.²⁸

bind at the same site and compete for the same binding site and compounds that can simultaneously bind to the protein to identify non-competitive binding sites (Fig. 5).

Dataset 3. Structure similarity/dissimilarity

The structural diversity of natural products can be analysed using radial fingerprints encoding the 2D topological atom environment (ECFP₄) using Canvas by Schrodinger (version 1.5.518).^{32,33} This type of fingerprint is one of the most widely used, and represents the environment of atoms in the neighbourhood of each heavy atom in the molecule within a

four-bond diameter. The total diversity is then represented by training a 25×25 self-organizing map (SOM) (Fig. 4C).³⁴ Each cell represents a cluster of compounds and the distance between cells (*i.e.* nearby cells are structurally related compounds) is indicated by the shading of the cell borders; darker borders indicate larger distance. Cells are coloured by population, with white for empty cells, and red for cell containing more than 5 compounds. The trained SOM is characterized by a toroidal architecture, which means that the top edge is connected to the lower edge and the left edge with the right edge.^{35–37}

Fig. 6 shows a seed compound (39) and other similar compounds in the cell. Compounds in the four adjacent cells would provide further data input.

Dataset 4. Scaffolds

To depict a molecule's complexity as completely as possible and remove ambiguities a scaffold networks approach uses the algorithm of Schuffenhauer *et al.* (Fig. 4D).³⁸ The approach decomposes a complex scaffold in every possible way, which then results in a network for each single molecule, consisting of scaffolds of different ring sizes.³⁹ This allows the exploration of the full scaffold space, such networks can be overlaid and hubs indicate similar properties.³⁹ The authors concluded, that the scaffold network approach should be utilized for the analysis of bioactivity and the alternate scaffold tree approach for the analysis of chemistry within libraries. A software solution is given through scaffold network generator, an open-source command-line utility, which can handle huge datasets.⁴⁰

For an illustration of this approach, the scaffold network of varioxepine A (40),⁴¹ yields 37 different embedded scaffolds; 5 scaffolds containing 5 rings, 9 scaffolds containing 4 rings, 9 scaffolds containing 3 rings, 7 scaffolds containing 2 rings and 7 consisting of single rings (Fig. 7).^{42,43} Natural product embedded scaffolds are often 3-dimensional and different to the concentration on flat molecules in other scaffolds used for medicinal chemistry.⁴⁴

Dataset 5. Molecular targets from biosynthetic enzymes

Recently, we have demonstrated that site comparison methodology using flavonoid biosynthetic enzymes as the query could automatically identify structural features common to different flavonoid-binding proteins, allowing the identification of flavonoid targets such as protein kinases.^{45,46} With the aim of further validating the hypothesis that biosynthetic enzymes and therapeutic targets can contain a similar natural product imprint, we collected a set of 159 X-ray structures representing 38 natural product biosynthetic enzymes by searching the Protein Data Bank. Each enzyme structure was used as a query to screen a repository of approximately 10 000 ligandable sites by active site similarity. We reported a full analysis of the screening results and highlight three retrospective examples where the natural product validates the method, thereby revealing novel structural relationships

Fig. 4 Proposed AI data integration to predict the function of a natural product based on its chemical structure A. Native MS observation of protein-ligand complexes. B. Network visualization of fragment hits against protein targets. C. Typical SOM of compound diversity. D. 2-ring scaffolds of varioxepine A. E. Natural products direct substituents in 3D space. F. Phenotypic response, hierarchically clustered based on their pairwise uncentred correlation coefficients.

between natural product biosynthetic enzymes and putative protein targets of the natural product. Natural product scaffolds direct substituents to interact with proteins (Fig. 4E).⁴⁷ Analysis of biosynthetic enzymes and target proteins, from a prospective view, has provided a list of up to 64 potential

novel targets for 25 well characterized natural products. As an example, pentalene (41) produced by the biosynthetic enzyme pentalene synthase correlates with serine/threonine-protein kinase Chk1, chorismate synthase, interleukin-2 (Fig. 8).⁴⁷

Partial data for datasets 1–5 is given in Table 1.

Fig. 5 Cartoon of a fragment-based network consisting of two proteins (A and B). Ligand 1 binds to both proteins indicating there is a similar localised binding site. Ligands 1, 2, 3 and 4 bind to protein B. Ligand 2 and ligand 3 bind to the same site and are competitive, the native MS would display 2 protein-ligand complexes. Ligand 2, and ligand 4 bind simultaneously and are non-competitive, the native MS shows the 2 individual protein-ligand complexes and a third protein-ligand complex due to both ligand 2 and ligand 4 simultaneously occupying the protein cavity. A similar situation arises for ligand 1 and ligand 4.

Data set 6. Phenotypic cellular activity

Unbiased phenotypic analysis has been used to examine natural products for a range of cellular effects (Fig. 4F). Cluster analysis was used to group similar responses. The data included compounds with known mechanism of action so that

Fig. 6 Compound diversity analysis to cluster compounds.

Fig. 7 Embedded two-ring scaffolds arising from a scaffolds network analysis of varioxepine A (40).

phenotypic data may subsequently lead to the discovery of particular targets.^{27,48,49} We used the cytological parameters followed by hierarchical clustering and visualisation as a dendrogram.^{27,48,49} AI can treat each data point separately.

Data for ajmaline (42), protopine (43) and gitoxigenin (44) is given in Table 2.

Artificial intelligence

Technological advances in artificial intelligence (AI), especially in the field of deep learning, hold the potential to make smart predictions based on explainable knowledge and patterns. Multi-faceted Big Data on the function of natural products offers exciting new opportunities to apply state-of-the-art deep learning advances to pursuing the grand challenge of predicting biological function from the chemical structure of a natural product. AI may tease out meaningful patterns and useful knowledge leading to integrated relationships and logic links between the metabolite structures and functions.

However, significant ground-work in applying this research is still critically needed before real breakthroughs can happen. For example, one open question in AI is how to explain deep learning outcomes from various heterogeneous data sources. Data integration facilitates cross-dataset analyses.⁵⁰ The opportunity is high given the vast quantity and variety of data.⁵⁰ There may be a new paradigm for data integration based on Representation Learning developments in

Artificial Intelligence, such as Autoencoders, that will provide a new paradigm for data integration.⁵⁰

The growth of metabolite Big Data creates an opportunity to learn patterns (extracting knowledge) from physical and chemical properties as well as biochemical function of known metabolites. The challenge is in designing and developing sophisticated machine learning algorithms that learn from the known data and generate robust prediction for the function of novel molecules.

The datasets on metabolomes with known ligand-protein binding pairs can be used to train deep learning neural networks to predict new ligand-protein binding pairs. In such a prediction process, machine learning challenges are in developing algorithms for: the recognition of binding pockets in a target, the pocket similarity between different targets, and matching molecules to identified pockets. Successful outcomes of these challenges will lead to the development of software tools for deciphering and predicting protein-metabolite interactions. On the other hand, genomic and proteomic approaches provide hypothesis driven information on molecular targets that can be analysed for pathways. They result, uniformly, in many potential targets.

AI can address two fundamental aspects. (1) *Phenotype Prediction*: the data creates an opportunity to learn models and mine patterns (extracting knowledge) from the physical and chemical properties and the phenotype function of known metabolites through deep learning and pattern analysis. (2) *Metabolite-Protein Interactions*: the large datasets currently available on metabolomes with known ligand-protein binding pairs can be used to train deep learning neural networks to predict new ligand-protein binding pairs. The research questions include: how to collect the training data and label the data with correct information? Could a deep learning-based model be developed to predict the specific targets of known metabolites? How can computational models be developed for verifying functions of metabolites? How to define the coupling relationships (*via* similarity or distance) between different chemical structures? How to evaluate the performance of prediction to reinforce the learning? How to tune the deep learning-based parameters? How to deal with the heterogeneous data sources? How to quantify the data interdependence of chemical structures and metabolite functions? How to build the AI bridge between metabolite structures and functions to make smart predictions?

Native MS allows specific interactions between metabolites and protein targets.^{15,28,29} Native MS can provide both the data analysis and the tool to confirm the specific ligand-protein complex *i.e.* the prediction of function to probe the AI predictions.

Conclusions

Is it time? Can AI use a matrix that contains information containing Smiles and PDB Codes and correlations embedded in six starting datasets to meaningful predict function (*i.e.* the protein target) of natural products.

Fig. 8 Structures of pentalene (41), ajmaline (42), protopine (43) and gitoxigenin (44).

Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1.	2. Fragment	4. Non-	6. 4	11. 1-ring
		5. Similar	7. 5-ring	10. 2-ring
			8. 4-ring	9. 3-ring
				13. Predicted

[illegible]

Table 2 Partial data for dataset 6

Compound	14. Nucleus area (μm^2)	15. Nucleus width (μm)	16. Nucleus length (μm)	17. Nucleus morphology ratio width to length	18. Nucleus roundness	19. Nucleus marker intensity	20. Nucleus texture index (SER spot 1 px)	21. Cell area (μm^2)	22. Cell width (μm)	23. Cell length (μm)
42	0.2	0.0	-0.5	0.6	0.6	0.0	0.5	-0.2	-0.3	0.3
43	0.0	0.2	0.0	-0.5	-0.1	0.0	1.8	-1.4	-1.0	-0.8
44	-1.0	-1.0	-0.7	-1.0	-1.5	0.3	2.5	-1.5	-1.0	-1.4

Compound	24. Cell morphology ratio width to length	25. Cell roundness	26. Tubulin marker intensity in cytoplasm	27. Tubulin marker intensity in outer region	28. Tubulin marker intensity in inner region	29. Tubulin texture index (SER spot 1px)	30. Mitochondria marker intensity in cytoplasm	31. Mitochondria marker intensity in outer region	32. Mitochondria marker intensity in inner region
42	0.0	0.0	0.3	-0.2	0.0	-0.7	-0.2	0.2	-0.2
43	0.2	0.6	1.6	0.1	1.5	-0.4	-0.1	0.6	-0.1
44	1.4	1.5	2.9	0.3	2.7	-1.1	0.2	1.7	0.2

Compound	33. Mitochondria marker texture index (SER spot 1px)	34. LC3b marker intensity in cytoplasm	35. LC3b marker intensity in outer region	36. LC3b marker intensity in inner region	37. LC3b marker texture spot 1px	38. Lysosomes marker intensity in cytoplasm	39. Lysosome marker intensity in outer region	40. Lysosomes marker intensity in inner region	41. Lysosomes marker texture index (SER spot 1 px)
42	-0.3	0.0	-0.2	0.0	-0.8	1.1	1.1	1.0	-0.9
43	0.4	1.0	-0.1	1.0	-1.0	1.5	1.4	1.4	0.2
44	-1.2	2.0	0.1	1.8	0.6	1.2	1.1	1.2	-0.4

Compound	42. Relative EEA1 marker spot signal in cytoplasm	43. Relative EEA1 marker spot signal in outer region	44. Relative EEA1 marker spot signal in inner region	45. EEA1 marker intensity in cytoplasm	46. Number of EEA1 spots per area of cytoplasm	47. Number of EEA1 spots in outer region	48. Number of EEA1 spots per area of outer region	49. Number of EEA1 spots in inner region	50. Number of EEA1 spots per area of inner region	51. EEA1 marker texture (SER spot 1 px)
42	0.7	0.4	0.7	-0.6	-0.5	-0.5	-0.3	-0.4	-0.2	0.0
43	1.1	1.0	0.8	-0.9	0.0	-0.8	0.0	-0.6	0.0	0.3
44	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The predicative value of machine learning for deep extraction of connectivities embedded in genomics and proteomics data is well recognised and utilised.⁵¹ For linear information molecules such as nucleic acids and proteins, the primary sequence structure can be readily digitised for computation and AI algorithms. For nonlinear information molecules such as natural products that are not amenable to linear coding, various manifolds of data sets have to be first collected, cross annotated, assembled and then multiplexed to capture, as much as possible, about these metabolites and their interactive partners, before the next set of AI algorithms can be implemented. While the scale and complexity of this deep learning may be unprecedented and daunting, it is the next grand challenge that will push functional chemogenomics and chemoproteomics into a new future. More importantly, the ability to predict function from structure for molecules such as natural products will forge a new paradigm for finding the next generations of precision medicines.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge the support of the Australian Research Council Discovery grant DP160101429.

Notes and references

- J. Neve, P. d. A. Leone, A. R. Carroll, R. W. Moni, N. J. Paczkowski, G. Pierens, P. Björquist, J. Deinum, J. Ehnebom, T. Inghardt, G. Guymer, P. Grimshaw and R. J. Quinn, *J. Nat. Prod.*, 1999, **62**, 324–326.
- J. A. Kalaitzis, P. d. A. Leone, L. Harris, M. S. Butler, A. Ngo, J. N. A. Hooper and R. J. Quinn, *J. Org. Chem.*, 1999, **64**, 5571–5574.
- M. S. Buchanan, A. Edser, G. King, J. Whitmore and R. J. Quinn, *J. Nat. Prod.*, 2001, **64**, 300–303.
- A. R. Carroll, G. K. Pierens, G. Fechner, P. de Leone, A. Ngo, M. Simpson, E. Hyde, J. N. A. Hooper, S.-L. Bostroem, D. Musil and R. J. Quinn, *J. Am. Chem. Soc.*, 2002, **124**, 13340–13341.
- A. R. Carroll, M. S. Buchanan, A. Edser, E. Hyde, M. Simpson and R. J. Quinn, *J. Nat. Prod.*, 2004, **67**, 1291–1294.
- A. R. Carroll, A. Ngo, R. J. Quinn, J. Redburn and J. N. A. Hooper, *J. Nat. Prod.*, 2005, **68**, 804–806.
- S. J. Rochfort, L. Towerzey, A. R. Carroll, G. King, A. Michael, G. Pierens, T. Rali, J. Redburn, J. Whitmore and R. J. Quinn, *J. Nat. Prod.*, 2005, **68**, 1080–1082.
- A. R. Carroll, G. Arumugan, R. J. Quinn, J. Redburn, G. Guymer and P. Grimshaw, *J. Org. Chem.*, 2005, **70**, 1889–1892.
- R. A. Davis, A. R. Carroll, S. Duffy, V. M. Avery, G. P. Guymer, P. I. Forster and R. J. Quinn, *J. Nat. Prod.*, 2007, **70**, 1118–1121.
- M. S. Buchanan, A. R. Carroll, R. Addepalli, V. M. Avery, J. N. A. Hooper and R. J. Quinn, *J. Org. Chem.*, 2007, **72**, 2309–2317.
- M. S. Buchanan, A. R. Carroll, G. A. Fechner, A. Boyle, M. M. Simpson, R. Addepalli, V. M. Avery, J. N. A. Hooper, N. Su, H. Chenc and R. J. Quinn, *Bioorg. Med. Chem. Lett.*, 2007, **17**, 6860–6863.
- M. S. Buchanan, A. R. Carroll, A. Edser, M. Sykes, G. A. Fechner, P. I. Forster, G. P. Guymer and R. J. Quinn, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 1495–1497.
- P. d. A. Leone, A. R. Carroll, L. Towerzey, G. King, B. M. McArdle, G. Kern, S. Fisher, J. N. A. Hooper and R. J. Quinn, *Org. Lett.*, 2008, **10**, 2585–2588.
- M. S. Buchanan, A. R. Carroll, D. Wessling, M. Jobling, V. M. Avery, R. A. Davis, Y. Feng, Y. Xue, L. Oster, T. Fex, J. Deinum, J. N. A. Hooper and R. J. Quinn, *J. Med. Chem.*, 2008, **51**, 3583–3587.
- H. Vu and R. J. Quinn, *J. Biomol. Screening*, 2008, **13**, 265–275.
- C. E. Zetterström, J. Hasselgren, O. Salin, R. A. Davis, R. J. Quinn, C. Sundin and M. Elofsson, *PLoS One*, 2013, **8**, e81969.
- J. E. Neve, H. P. Wijesekera, S. Duffy, I. D. Jenkins, J. Ripper, S. J. Teague, A. Garavelas, G. Nikolakopoulos, P. V. Le, P. d. A. Leone, N. B. Pham, P. Shelton, N. Fraser, A. R. Carroll, V. M. Avery, C. McCrae, N. Williams and R. J. Quinn, *J. Med. Chem.*, 2014, **57**, 1252–1275.
- T. Grkovic, R. H. Pouwer, Q. Wang, G. P. Guymer, J. Holst and R. J. Quinn, *J. Nat. Prod.*, 2015, **78**, 1215–1220.
- L.-W. Tian, Y. Feng, T. D. Tran, Y. Shimizu, T. Pfeifer, H. T. Vu and R. J. Quinn, *Bioorg. Med. Chem. Lett.*, 2017, **27**, 4007–4010.
- R. J. Quinn, R. P. Gregson, A. F. Cook and R. T. Bartlett, *Tetrahedron Lett.*, 1980, **21**, 567–568.
- L. P. Davies, K. M. Taylor, R. P. Gregson and R. J. Quinn, *Life Sci.*, 1980, **26**, 1079–1088.
- A. F. Cook, R. T. Bartlett, R. P. Gregson and R. J. Quinn, *J. Org. Chem.*, 1980, **45**, 4020–4025.
- M. J. Dooley and R. J. Quinn, *J. Med. Chem.*, 1992, **35**, 211–216.
- M. J. Dooley and R. J. Quinn, *Bioorg. Med. Chem. Lett.*, 1992, **2**, 1199–1200.
- S. Urban, P. d. A. Leone, A. R. Carroll, G. A. Fechner, J. Smith, J. N. A. Hooper and R. J. Quinn, *J. Org. Chem.*, 1999, **64**, 731–735.
- T. Grkovic, R. H. Pouwer, M. L. Vial, L. Gambini, A. Noël, J. N. Hooper, S. A. Wood, G. D. Mellick and R. J. Quinn, *Angew. Chem., Int. Ed.*, 2014, **53**, 6070–6074.
- D. Wang, Y. Feng, M. Murtaza, S. Wood, G. Mellick, J. N. A. Hooper and R. J. Quinn, *J. Nat. Prod.*, 2016, **79**, 353–361.
- H. Vu, L. Pedro, T. Mak, B. McCormick, J. Rowley, M. Liu, A. D. Capua, B. Williams-Noonan, N. B. Pham, R. Pouwer, B. Nguyen, K. T. Andrews, T. Skinner-Adams, J. Kim, W. Hol, R. Hui, G. J. Crowther, W. C. V. Voorhis and R. J. Quinn, *ACS Infect. Dis.*, 2018, **4**, 431–444.
- L. Pedro and R. J. Quinn, *Molecules*, 2016, **21**, 984.
- H. Vu, C. Roullier, M. Campitelli, K. R. Trenholme, D. L. Gardiner, K. T. Andrews, T. Skinner-Adams, G. J. Crowther, W. C. Van Voorhis and R. J. Quinn, *ACS Chem. Biol.*, 2013, **8**, 2654–2659.

- 31 P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
- 32 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
- 33 R. Glen, A. Bender, C. Arnby, L. Carlsson, S. Boyer and J. Smith, *IDrugs*, 2006, **9**, 199–204.
- 34 T. Kohonen, *Biol. Cybern.*, 1982, **43**, 59–69.
- 35 M. Pascolutti, M. Campitelli, B. Nguyen, N. Pham, A.-D. Gorse and R. J. Quinn, *PLoS One*, 2015, **10**, e0120942.
- 36 M. Pascolutti and R. J. Quinn, *Drug Discovery Today*, 2014, **19**, 215–221.
- 37 Y. Feng, M. Campitelli, R. A. Davis and R. J. Quinn, *Mar. Drugs*, 2014, **12**, 1169–1184.
- 38 A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch and H. Waldmann, *J. Chem. Inf. Model.*, 2007, **47**, 47–58.
- 39 T. Varin, A. Schuffenhauer, P. Ertl and S. Renner, *J. Chem. Inf. Model.*, 2011, **51**, 1528–1538.
- 40 M. Matlock, J. Zaretski and S. Swamidass, *Bioinformatics*, 2013, **29**, 2655–2656.
- 41 P. Zhang, A. Mándi, X.-M. Li, F.-Y. Du, J.-N. Wang, X. Li, T. Kurtán and B.-G. Wang, *Org. Lett.*, 2014, **16**, 4834–4837.
- 42 S. Böttcher, A. Di Capua, J. W. Blunt and R. J. Quinn, in *Blue Biotechnology: Production and use of marine molecules*, ed. S. La Barre and S. S. Bates, Wiley-VCH Verlag GmbH & Co., Weinheim, Germany, 2018, vol. 1, pp. 297–321.
- 43 F. M. Tajabadi, M. R. Campitelli and R. J. Quinn, *Springer Science Reviews*, 2013, vol. 1, pp. 141–151.
- 44 B. Zdrazil and R. Guha, *J. Med. Chem.*, 2018, **61**, 4688–4703.
- 45 B. M. McArdle, M. R. Campitelli and R. J. Quinn, *J. Nat. Prod.*, 2006, **69**, 14–17.
- 46 E. Kellenberger, A. Hofmann and R. J. Quinn, *Nat. Prod. Rep.*, 2011, **28**, 1483–1492.
- 47 N. Sturm, R. J. Quinn and E. Kellenberger, *Planta Med.*, 2018, **84**, 304–310.
- 48 M.-L. Vial, D. Zencak, T. Grkovic, A.-D. Gorse, A. Mackay-Sim, G. D. Mellick, S. A. Wood and R. J. Quinn, *J. Nat. Prod.*, 2016, **79**, 1982–1989.
- 49 Y. Dashti, M. L. Vial, S. A. Wood, G. D. Mellick, C. Roullier and R. J. Quinn, *Tetrahedron*, 2015, **71**, 7879–7884.
- 50 V. Vijayan, A. D. Rouillard, D. K. Rajpal and P. Agarwal, *Expert Opin. Drug Discovery*, 2019, 1–4.
- 51 D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello and J. J. Collins, *Cell*, 2018, **173**, 1581–1592.