Research article

NPClassifier: A deep neural network-based structural classification tool for natural products

Hyun Woo Kim¹, Mingxun Wang^{2,3}, Christopher A. Leber¹, Louis-Félix Nothias², Raphael Reher¹, Kyo Bin Kang⁴, Justin J. J. van der Hooft⁵, Pieter C. Dorrestein², William H. Gerwick,^{1,2*} and Garrison W. Cottrell^{6*}

Affiliation

¹Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego La Jolla, CA, USA.

²Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA.

³Ometa Labs LLC, San Diego, CA, USA.

⁴College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea

⁵Bioinformatics Group, Wageningen University, Wageningen, The Netherlands.

⁶Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA.

*Joint senior authors

Abstract

Computational approaches such as genome and metabolome mining are becoming essential to natural products (NP) research. Consequently, demands for automated NP classification system for massive data are increasing. The semantic ontology of NPs classifies molecules based on the taxonomy of the producing organism, the nature of the biosynthetic pathway, their biological properties, as well as the presence of chemical substructures. Thus, a holistic and automatic NP classification framework could have considerable value to comprehensively navigate the relatedness of NPs. Here, we introduce NPClassifier, the first deep-learning tool for the automated structural classification of NPs. We expect that NPClassifier will accelerate NP discovery by facilitating and enabling large-scale genome and metabolome mining efforts and linking of NP structures to their underlying bioactivity.

Introduction

‘Classification’ is a systematic arrangement of elements into groups or categories according to established criteria to recognize, differentiate and understand ideas or objects. In natural product (NP) or specialized metabolite-guided drug discovery, NPs are categorized based upon their molecular structures, chemical properties, bioactivities, and biosynthetic pathways. NPs are an essential resource for drug design and discovery, as well as for pharmacological tools used in biomedical applications.¹⁻³ Fundamentally, molecules belonging to the same class share similar properties based on the criteria used in the classification scheme; therefore, the classification facilitates exploration of large regions of chemical space so as to derive useful information, such as new sources of drugs or bioactivity profiles.^{4,5}

Existing ontologies that provide chemical structure-based classifications include the ChEBI ontology,⁶ the Medical Subject Heading (MeSH) thesaurus with PubChem,⁷ LIPID MAPS,⁸ and NP-specific databases such as Super Natural II,⁹ MIBiG¹⁰, the Natural Products Atlas,¹¹ and the Dictionary of Natural Products (<http://dnp.chemnetbase.com>). The molecules in these databases are curated by structural classes, biological activities, or source organisms. These structures and their classification terms are used to train various tools for NP research.¹²⁻¹⁴ The ontologies and structures in these databases were manually curated from the literature. However, because most of these resources are not open for community-based curation, they are static and incomplete, and because they are necessarily restricted to the molecules included in their databases, they cannot be used to automatically classify molecules. To tackle this last challenge, ClassyFire was developed to automatically classify molecular structure based on chemical properties into the ChemOnt ontology, a well-defined chemical hierarchy.¹⁵ Nevertheless, ClassyFire was designed for general organic and bio-organic chemistry, primarily aimed at the metabolomics and exposomic communities, and only provides partial

classifications with semantic knowledge from NPs, thus significantly reducing its relevance for NP research.

Unlike structure-only based classifications, traditional classification of NPs encompasses structural information as semantically defined by NP researchers since the 1800s. For example, cyanopeptolins are a typical class of cyclic depsipeptide isolated from cyanobacteria.¹⁶ This class typically contains a (3*S*)-amino-(6*R*)-hydroxy piperidone (AHP)-moiety and is produced via a non-ribosomal peptide synthetase pathway. As they are well-known to display anti-cancer bioactivity, when a novel cyanopeptolin derivative is discovered, it can be hypothesized that this compound also has bioactivity against cancer cells.¹⁷ Knowledge concerning the characteristics and properties of NP classes are continuously expanded and revised with novel discoveries made by NP researchers. This ensures that, over time, NP classifications are semantically largely consistent and informative. Consequently, the established classification ontology for NPs allows a broader understanding of NPs.

NPs have a prolific structural diversity. In part, this is the result of the large number of possible biosynthetic mechanisms including hybridization and rearrangement of partial structures,¹⁸ and can also involve multiple organisms or inorganic reactions taking place due to the environment. To understand and classify NPs, various rule-based approaches such as analyzing functional groups, comparing structure similarity, and finding maximum common substructures (MCS) have been attempted.¹⁹⁻²¹ However, a rule-based system is inherently limited in producing a comprehensive and relevant classification because: 1) rules always have exceptions; 2) they might not encompass the entire space of compounds; and 3) structural features can be shared by multiple categories, which makes results difficult to interpret.²² For example, the rule-based definition of ‘limonoids’ is any triterpenoid that is highly oxygenated and has a prototypical structure either containing or derived from a precursor with a 4,4,8-

trimethyl-17-furanylsteroid skeleton.²³ Nevertheless, some highly modified limonoids or limonoid-like compounds are not consistent with this structure definition.^{24,25} As a result, current rule-based structural classification tools or ontologies have limited use for NP classification.

To develop a new classification tool for automated classification of NPs with traditional knowledge, we focused on the consistency of the traditional NPs classification system. Over the last two decades, an average of 1,600 new marine and microbial NPs have been reported annually,²⁰ and most were reported with their NP classifications during the peer-review and publishing process. This provides the consistency and sustainability for classification of NPs based on the contribution of NPs community. Thus, the dataset of NPClassifier was mainly established based on standard practices in the literature concerning chemical entities and their classification.

As data size increases in NP research, deep neural networks (DNNs) have risen as an alternative solution for enhancing drug discovery, genome mining, and structure elucidation.^{13, 26-29} The power of deep learning is largely dependent on how the features are extracted from the data. This is because in contrast to traditional machine learning approaches or rule-based classifications, DNNs learn features from the data via back propagation during training.³⁰ Therefore, DNNs are an attractive method to mimic the basis by which knowledge experts have applied traditional classifications to NPs.

In this paper, we introduce a deep neural network-based NP classification tool called ‘NPClassifier’, which is freely available at <https://npclassifier.ucsd.edu> together with a web-API (see Supporting Information). NPClassifier was developed using supervised feed-forward networks with 73,607 NPs collected from public databases including Pubchem, ChEBI, Chempid, and the Universal Natural Products Database (UNPD).³¹⁻³⁴ NPClassifier classifies

the structure of an NP at three levels; these are organized into 7 Pathways, 70 Superclasses, and 653 Classes, all of which are generally recognized by the NP research community (Figure 1).

(B)

NP Classifier
Version - Release_1

CC1(C)C(C=CC2(C)C1=C(C(=O)C(C3(C)C2CCC4(C)C35C(CC4C6=COC=C6)O5)=O)=O

O=C1C(C)C(C=CC2(C)C1=C(C(=O)C(C3(C)C2CCC4(C)C35C(CC4C6=COC=C6)O5)=O)=O

entry	type
Limonoids	class
Triterpenoids	superclass
Terpenoids	pathway

JSON format

```
{
  "class_results": ["Limonoids"],
  "superclass_results": ["Triterpenoids"],
  "pathway_results": ["Terpenoids"],
  "isglycoside": false,
}
```

Figure 1. Overview of NPClassifier. (A) In the data preparation stage, compound names (synonyms) and their class information were collected from the literature. The compound names were converted to chemical fingerprints and class information was assigned based on the NPClassifier ontology. During the training phase, molecular fingerprints were input to a deep neural network. The binary cross-entropy loss was calculated by comparing the prediction result from sigmoid output and ground truth, and back propagated to adjust the model parameters. In classification, a submitted chemical structure is classified with three levels of Pathway, Superclass, and Class by NPClassifier. (B) The webpage of NPClassifier and JSON-formatted output for massive inputs.

Results and Discussion

Training, optimization, and evaluation of NPClassifier models.

Classification system. A classification system was established based on the literature from the specialized metabolism of plants, marine organisms, fungi, and microorganisms^{16, 35-37} The MIBiG database,¹⁰ which provides BGCs (biosynthetic gene clusters) information of NPs was used to ensure the correctness of the biosynthetic pathways for each class. The categories used in NPClassifier are defined at three hierarchical levels: Pathway, Superclass, and Class.

The Pathways of NPClassifier consist of seven categories: fatty acids, polyketides, shikimates-phenylpropanoids, terpenoids, alkaloids, amino acids/peptides, and carbohydrates. The fatty acids and polyketides are major biosynthetic pathways of microorganisms and relate to the production of many antibiotics (e.g., doxycycline, erythromycin, and azithromycin) or immunosuppressants (e.g., Tacrolimus, Rapamycin). The shikimates-phenylpropanoids category is based on the shikimate pathway. The phenylpropanoids are a diverse family of organic compounds from the shikimate pathway. Aromatic amino acids and many aromatic NPs are formed from the phenylpropanoids via the shikimate pathway.³⁸ The terpenoids are a large and diverse category of NPs derived from the mevalonate (MVA) or the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway. Terpenoids have diverse biological properties, including cytotoxicity and anti-inflammatory effects.³⁹ Alkaloids represent nitrogenous organic compounds from NPs without obvious amino acids or other peptidic characteristics. Many alkaloids are part of traditional medicine or have found use as single molecule drug candidates with their unique bioactivity attributed to the presence of basic nitrogen atoms.⁴⁰ The amino acid/peptides category is related to different biochemical mechanisms for peptide synthesis, wherein multiple amino acids are linked via amide (or peptide) bonds. The ribosomal or nonribosomal peptide synthetase biosynthetic machinery is responsible for the formation of

this category of NP and has been widely investigated with genome sequencing approaches. The carbohydrate category represents specialized metabolites with biological activity that mainly consist of carbohydrate substructures.

The Superclasses represent sub-categories within the Pathways, and at the present time 70 designations are proposed. The categories in the Superclass originate from the general classes of metabolites (ex. flavonoids, meroterpenoids, or steroids), the general chemical or molecular shapes (e.g., chromanes, phloroglucinols, or macrolides), or from biosynthetic information (e.g., tryptophan alkaloids, aromatic polyketides, or pseudo alkaloids). The chemical properties or taxonomic information of the chemical entities can be expected to be associated with the Superclass. For example, one of the Superclasses, steroids, is a well-known biologically active metabolite group with a specific multi-ring architecture and consistent chemical properties.⁴¹

The Superclasses are subdivided into Classes, that represent specific compound families (e.g., erythromycins, penicillins, or cannabinoids), characteristic functional groups (e.g., chromones, azaphilones, indole alkaloids, or 3-spiro tetramic acids), or scaffold diversity within Superclasses (e.g., flavans, flavones, and chalcones from flavonoids). In this manner, 653 Classes were identified in the NPClassifier. (see Supporting Information for a complete list of Pathways, Superclasses and Classes).

Additionally, glycosides are also detected by NPClassifier. A glycoside is any molecule in which one or more sugar groups are bonded through glycosidic bonds between anomeric carbon and non-sugar parts. Regarding its numerous important roles in NPs, distinguishing between glycosides and their aglycone is essential to understand NPs.⁴² The result of glycoside detection is provided together with the three-level classification system. (see Supporting Information for additional details about detecting glycosides)

Figure 2. An example of the classification ontology of NPClassifier. (A) The amino acids-peptides Pathway and its Superclasses and Classes in the NPClassifier classification system. This Pathway contains 12 Superclasses and 51 Classes. (B) The macrolides Superclass is involved in both polyketides and amino acids-peptides Pathways. (C) The peptide alkaloids Superclass and its Classes belong to both alkaloids and amino acids-peptides Pathways

Figure 2A showed the classification system for the amino acids-peptides Pathway, in which 12 Superclasses and 51 Classes were included. Among the Superclasses of the amino acids-peptides Pathway, some Superclasses such as macrolides and peptide alkaloids are included in multiple categories. As shown in Figure 2B, the antimycin Class, which is included in the macrolides Superclass, biosynthetically belongs to hybrid NRPS-PKS synthases. Consequently, this Class is shared between both the amino acids-peptides and polyketides Pathways. The peptide alkaloids are generally considered as alkaloids but are also composed of natural amino acids linked by amide bonds.⁴³ Thus, the peptide alkaloids Superclass is included in both the amino acid-peptides and alkaloids Pathways; hence, the classification system in the

NPClassifier has an acyclic graph structure to consider the hybrid system of NPs.

Chemical Descriptors. The Morgan fingerprint method was chosen as the format for inputting structural information to the neural network. Morgan fingerprints were designed to encode the structure of a molecule in a form that allows for rapid and efficient quantification of molecular structure similarity or finding matches to a query substructure. This binary molecular fingerprint method can represent the presence (1) and absence (0) of the atoms or substructures in molecules, but cannot represent the number of atoms or substructures in the molecules, which is an important factor for NP classification. Thus, the binary Morgan fingerprint method was modified to an integer method to include the count of atoms and substructures using RDKit version 2019.09.3 (Figure 3).⁴⁴ In our modified Morgan fingerprint method (MMF), the entries in the vector have non-negative integer values that allow us to provide the network with the count of each type of atom and substructures in the molecular structures. This information is particularly important for the classification of oligomeric or structurally iterative NPs such as proanthocyanidins, tannins or carotenoids.

Figure 3. The chemical descriptor and the deep learning architecture of NPClassifier (A) The difference between Morgan fingerprints and modified Morgan fingerprints, the latter were developed for this application. Morgan fingerprints are presented in a binary data format over all radii. Alternatively, the modified Morgan fingerprints have an integer format (B) The top layers for classification were different, depending on the classification level as indicated in the legend.

Model Optimization and evaluation. During the training experiments, the performance of the models using two kinds of chemical descriptors was compared at the Class level; Morgan fingerprints (MFs) and our modified-Morgan fingerprints (MMFs) as described above. As shown in Table 1, the loss of the MMFs-based model was significantly lower, and the cosine similarity and mean average precision (mAP) of the MMFs-based model were significantly

higher than that of MFs-based model in Superclass and Class. Although the mean Average Precision was higher for the standard MFs, since we use Cosine similarity for retrieval of similar compounds, we used the MMFs for all three. Hence, MMFs were chosen as the input format for the DNN. These results also indicated that the semantic knowledge-based classification ontology for NPs was consistent with our expectations, resulting in excellent accuracy by the deep neural networks.

Table 1. Comparison of the losses, cosine similarities and mean average precisions (mAP) from the neural networks using different chemical descriptors.^a

Model	Classification levels	Loss (SD)	Cosine similarity (SD)	mAP (SD)
MFs (binary)	Pathway	0.0197 (0.0004)	0.9863 (0.0004)	0.9932 (0.0003) ^{**}
	Superclass	0.0050 (0.0000)	0.9642 (0.0003)	0.9423 (0.0010)
	Class	0.0012 (0.0000)	0.9314 (0.0002)	0.8734 (0.0018)
MMFs (Integer)	Pathway	0.0211 (0.0016)	0.9849 (0.0013)	0.9920 (0.0004)
	Superclass	0.0046 (0.0001) ^{**}	0.9682 (0.0009) ^{**}	0.9515 (0.0030) ^{**}
	Class	0.0010 (0.0000) ^{***}	0.9377 (0.0005) ^{***}	0.8951 (0.0022) ^{***}

^aEach model was optimized based on results from the validation set (n=11,777) and evaluated by using the test set (n=14,721). The results are the average values over five runs of each model. There was no significant difference in the Pathway Loss or Cosine similarity between the two models.

* significant at $p < 0.05$; ** significant at $p < 0.005$; *** significant at $p < 0.001$.

SD = standard deviation. MFs = Morgan fingerprints. MMFs = Modified Morgan fingerprints.

For the three levels of classification of NPs, two different architectures using MMFs were compared; one with three separate single-task classifiers, and a multi-task model. In the multi-task model, the output layers were divided into three different layers to predict the three levels (Pathway, Superclass, and Class) simultaneously; this allowed for the hidden layers to receive feedback from the three output layers.⁴⁵ At the Pathway and Superclass levels, the loss and cosine similarity were improved by using the single-task classifiers. At the Class level, the results from the single-task model were significantly better in cosine similarity (0.0053 higher) and mAP (0.0238 higher) (Table 2); hence, the three single-task models were chosen as the models for NPClassifier. (see Supporting Information for additional details about the model optimization and metrics)

Table 2. Comparison of the losses, cosine similarities and mean average precisions (mAPs) from the multi-task, single-task and single-task with data augmentation models.^a

Model	Classification levels	Loss (SD)	Cosine similarity (SD)	mAP (SD)
Multi-task	Pathway	0.0234 (0.0006)	0.9694 (0.0009)	0.9928 (0.0004)
	Superclass	0.0049 (0.0001)	0.9571 (0.0006)	0.9551 (0.0016)
	Class	0.0011 (0.0000)	0.9324 (0.0008)	0.8713 (0.0009)
Single-task	Pathway	0.0211 (0.0016)*	0.9849 (0.0013)***	0.9920 (0.0004)
	Superclass	0.0046 (0.0001)*	0.9682 (0.0009)***	0.9515 (0.0030)
	Class	0.0010 (0.0000)	0.9377 (0.0005)***	0.8951 (0.0022)***

^aEach model was optimized based on the result from the validation set (n=11,777) and evaluated by using the test set (n=14,721). The results are the average values over five runs.

* significant at $p < 0.05$; ** significant at $p < 0.005$; *** significant at $p < 0.001$.

SD = standard deviation

Performance of the NPClassifier vs. Classyfire. To evaluate the general performance of NPClassifier, it was tested with an external test set. The test set contained representatives from three of the Pathways (amino acid-peptides, polyketides, and terpenoids) and Superclasses (flavonoids, steroids, and lignans), which were established from the Dictionary of Natural Products. These Pathways and Superclasses were chosen because they represent overlapping categories between NPClassifier and Classyfire¹⁵, allowing direct comparison. Because Classyfire is a deterministic system, we cannot report means and standard deviations, so the results are reported for one run for both classifiers. Each Pathway and Superclass had 1,000 chemical entities, resulting in 6,000 in total. We report precision (PRE), recall (REC), and F1 score (the harmonic mean of precision and recall) in Table 3. As can be seen from the Table 3, NPClassifier outperformed ClassyFire and generally showed excellent results. NPClassifier was especially outstanding in recognizing polyketides and lignans. The F1 scores showed that NPClassifier returned equal or better scores in the majority of cases.

The performance of NPClassifier and ClassyFire was investigated in more detail in NP annotation at the Class level. A total of 62 classes that, at minimum, contained 100 chemical entities were tested on both platforms, as shown in Figure 4. Again, NPClassifier outperformed

Classyfire for 50 Classes, and performed equally or slightly worse results for the remaining 12 Classes (see Supporting Information for additional details on the external test set).

Table 3. Comparison of the performance between NPClassifier and ClassyFire on the external test set from the Dictionary of Natural Products (n=6,000) ^a

Classification levels	Name	NPClassifier			ClassyFire		
		PRE	REC	F1 score	PRE	REC	F1 score
Pathway	Amino acids and peptides	0.965	0.850	0.904	0.965	0.860	0.909
	Polyketides	0.970	0.855	0.909	1.000	0.370	0.540
	Terpenoids	0.888	0.975	0.929	0.833	0.690	0.755
Superclass	Flavonoids	0.937	0.937	0.937	0.878	0.908	0.893
	Steroids	0.999	0.982	0.990	0.933	0.892	0.912
	Lignans	0.977	0.690	0.809	0.992	0.353	0.521

^a Each class had 1,000 chemical entities equally

Figure 4. Comparison of the classification results from NPClassifier (Blue) and ClassyFire (Orange). Overlap is Brown. Chemical entities (n=6,200, 100 chemical entities for each) from 62 classes were subjected to NPClassifier and ClassyFire and the classification accuracy was measured, respectively. NPClassifier showed better results for 42 classes, and equal or slightly worse results for 20 classes compared with ClassyFire.

Interpretation of models. DNNs are sometimes considered to be ‘Black boxes’ because it

can be difficult to explain how they work, and understanding the model is important to improving its performance. For example, why does it fail in some cases, and how can its reliability be improved? To this end, the change in response from NPClassifier was tracked as a function of modified molecular structures. This allowed us to evaluate whether the model classified a molecule based on class related structural moieties. As shown in Figure 5A, the ester bonds in the cyclic depsipeptide were replaced with amide bonds, and the classes of cyclic and depsipeptides were changed to cyclic peptides. In the same fashion, when the C ring of the flavonone was modified from an 4-ketone to an 3-hydroxy-4-ketone, the results were changed from flavonones to dihydroflavonols, and when the C ring of the dihydroflavonol was modified from an 3-hydroxy-4-ketone to 3-hydroxy, the results were changed from dihydroflavonols to flavan-3-ols (Figure 5B). From these experiments, we can observe that the classification results from NPClassifier are influenced by specific structural moieties following the general definition of the flavonoids and peptides. This indicates that the neural networks extracted the class related features from the molecular structures.

Additionally, the cases where NPClassifier failed were investigated to understand the conditions underlying these incorrect classifications. The ferulic acid trimers were classified as a diarylheptanoid; the correct result should have been shikimates and phenylpropanoids (Pathway), phenolic acids (Superclass) and cinnamic acids and derivatives (Class). Both ferulic acids and diarylheptanoids are from phenylpropanoids but the linkage patterns are different. The compound in Figure 5C is a rarely reported NP; searching the SciFinder database returned only 12 compounds with a similarity of over 85%. Thus, it is likely that NPClassifier does not have enough training data to properly classify this compound at the Superclass and Class levels.

Figure 5. Examples of the correlations between the structural modifications and classification results, and conflicting or missing classifications. (A) Ester bonds of a cyclic depsipeptide were replaced with amide bonds, and the classification changed from cyclic and depsipeptides to cyclic peptides. (B) The correlation between the modification of C ring substituents in flavonoids and the resulting classification. (C) Structure of ferulic acid trimer.

In summary, the incorrect output of NPCClassifier can be traced to a deficiency in the training dataset and class system. In other words, the addition of more entries to the reference dataset or expanding the class system is expected to solve such deficiencies. For those purposes, user-feedback and evaluation forms were created on the NPCClassifier website to collect community's contributions and/or suggestions.

Application of NPClassifier to natural products research and drug discovery.

NPClassifier was designed for natural products classification and can assist natural product research in a variety of ways. In the present case, we introduce the application of NPClassifier to the Natural Products Atlas (NPAtlas) database, as described below.

Analysis and interpretation of databases. The NPAtlas is an open source database that provides compound names, chemical structures, organism sources, and a structure similarity-based chemical space for natural products from fungi and bacteria along with their references.¹¹ Using NPClassifier, we can compare the differences in specialized metabolites produced by fungi and bacteria (Figure 6A and 6B). Both groups of organisms share an abundance of polyketide specialized metabolites, but macrolides and aromatic polyketides are major scaffolds for natural products originating from bacteria, whereas aromatic polyketides and naphthalenes have mainly been reported to originate from fungi. Such classification results can be applied to find producers of desired metabolite scaffolds from the NPAtlas database. For example, epothilones are macrolide compounds with a well-known potential as anticancer drugs.⁴⁶ Using the classification from NPClassifier, 45 epothilone derivatives were successfully found in the NPAtlas database, and *Sporangium* and *Myxococcus* species were reported as epothilone producers (see Supporting Information for all classified results of the NPAtlas database using NPClassifier).

Natural product scaffolds-based *in silico* screening. Over the past few decades, virtual screening has improved in terms of the size of real as well as virtual compound databases available for searching, and the algorithms applied to them are providing improved results.⁴⁷ Natural products possess high structural diversity, although this diversity is readily mapped to the outputs of specific biosynthetic pathways. If a specific scaffold is chosen as a candidate structure type from *in silico* screening, then using NPClassifier, a researcher can search for

source organisms that produce the targeted chemical class. This can be integrated with modern techniques of engineered biochemical pathways and synthetic biology.

Figure 6. Application of NPClassifier to natural products research and drug discovery. Distribution of structural diversity from bacteria (A) and fungi (B) with classification results. (C) Structure classification-based heat map for virtual screening results for blood stage malaria inhibition and (D) SARS-CoV2 3CL Mpro inhibition. The red color signifies predicted high potency and the blue color signifies predicted low potency (input data from NPAtlas).

For example, using the malaria inhibitor prediction (MAIP) platform from EMBL-EBI (<https://www.ebi.ac.uk/chembl/maip/>), 25,523 chemical entities from the NPAtlas were screened to find potential natural product-derived antimalarial agents. Then all chemical entities were classified by NPClassifier and assigned anti-malarial scores from the MAIP platform. As shown in Figure 6C, an obvious trend in the potential hit molecules was for the macrolide and oligopeptide Superclasses. In the macrolides Superclass, the ascomycins-

rapamycins Class shows the most potent activity with a 132.8 average model score from MAIP, similar to results from previous studies.⁴⁸ On the other hand, another macrolide Class, the zearalenones, did not show the predicted inhibitory effects. Among small molecules, the tryptophan alkaloids Superclass, including the carbazole alkaloids, ergot alkaloids, and penitrems Classes, and the peptide alkaloids Superclass, including the indole diketopiperazine alkaloids Class all showed potential as antimalarial agents. These results illustrate that using the classification results from NPClassifier improves understanding of what kinds of scaffolds are likely to show a desired biological activity.

Additionally, we used a virtually screened NPAtlas to identify potential inhibitors of the SARS-CoV-2 3CL main protease (Mpro) (<https://www.aicures.mit.edu/>).⁴⁹ By analysis of the correlation between predicted activities and the scaffolds classified by NPClassifier, terphenyls, naphthalenes and chromones showed good predicted activities. Interestingly, chromones have been studied as an early anti-inflammatory treatment for SARS-CoV-2 infection.^{50, 51} From the NPAtlas database classified by NPClassifier, many fungi are identified as producing chromone-type natural products via polyketide synthases. Thus, fungi might be an interesting reservoir in which to search for new chromone derivatives for SARS-CoV-2 treatment (Figure 6D).

Conclusion

In this study, we introduce NPClassifier, a tool that was designed for the classification of natural products using deep learning, along with a specific training strategy, optimized parameters, and its evaluation and application. The classification ontology used in the NPClassifier was categorized into three hierarchical levels based on semantic knowledge. The specialized metabolism (Pathway), chemical properties or chemotaxonomic information (Superclass), and structural details (Class) of the classified molecules can be deduced from the classification results. We anticipate that by supporting large-scale computationally driven NP

discovery studies, for example to link the results from genome and metabolome mining, NPClassifier will not only be used in antibiotics and drug discovery applications but also to understand molecular details in ecology and human health⁵².

Supporting information

The Supporting Information is available free of charge on the ACS Publications website at DOI: xx.xxxx/acscentsci.xxxxxxx.

- The introduction of NPClassifier website and web-API, experimental, and classification schema of NPClassifier (PDF)
- The classified results of the Natural Product Atlas database with the virtual screening data (xlsx)
- Dataset for training NPClassifier (xlsx)

The web application is available at <https://npclassifier.ucsd.edu/>

The source code for the NPClassifier is available on GitHub at <https://github.com/mwang87/NP-Classifier>.

Acknowledgments

We gratefully acknowledge financial support by: the U.S. National Institutes of Health (NIH) grant on tools for rapid and accurate structure elucidation of natural products (R01 GM107550) to P.C.D., W.H.G., and G.W.C., the Gordon and Betty Moore Foundation grant (GBMF7622) to P.C.D., G.W.C., and G.W.C., and the Netherlands eScience Center (NLeSC) ASDI eScience grant (ASDI.2017.030) to J.J.J.v.d.H.

References

1. Lachance, H.; Wetzel, S.; Kumar, K.; Waldmann, H. Charting, Navigating, and Populating Natural Product Chemical Space for Drug Discovery. *J. Med. Chem.* **2012**, *55* (13), 5989-6001.
2. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J. A.; Tagliabue, S. G.; Todeschini, R.; Schneider, G. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **2018**, *1* (1), 44.
3. Wu, M. C.; Law, B.; Wilkinson, B.; Micklefield, J. Bioengineering natural product biosynthetic pathways for therapeutic applications. *Curr. Opin. Biotechnol.* **2012**, *23* (6), 931-40.
4. Reymond, J. L.; Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *Acs. Chem. Neurosci.* **2012**, *3* (9), 649-657.
5. Saldivar-Gonzalez, F. I.; Lenci, E.; Trabocchi, A.; Medina-Franco, J. L. Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study. *Rsc. Adv.* **2019**, *9* (46), 27105-27116.
6. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic. Acids. Res.* **2016**, *44* (1), 1214-1219.
7. Rogers, F. B. Medical subject headings. *Bull. Med. Libr. Assoc.* **1963**, *51* (1), 114-6.
8. Fahy, E.; Subramaniam, S.; Murphy, R. C.; Nishijima, M.; Raetz, C. R. H.; Shimizu, T.; Spener, F.; van Meer, G.; Wakelam, M. J. O.; Dennis, E. A., Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid. Res.* **2009**, *50*, 9-14.
9. Banerjee, P.; Ereman, J.; Gohlke, B. O.; Wilhelm, T.; Preissner, R.; Dunkel, M., Super Natural II-a database of natural products. *Nucleic. Acids. Res.* **2015**, *43* (1), 935-939.
10. Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Munoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Duran, H. G. S.; Andreu, V. P., et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic. Acids. Res.* **2020**, *48* (1), 454-458.
11. van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castano-Espriu, L.; Chang, C.; Clark, T. N., et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *Acs. Central. Sci.* **2019**, *5* (11), 1824-1833.
12. Seo, M.; Shin, H. K.; Myung, Y.; Hwang, S.; No, K. T., Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP)

for natural product-based drug development. *J. Cheminformatics*. **2020**, *12* (1), DOI: 10.1186/s13321-020-0410-3

13. Dührkop, K.; Nothias, L. F.; Fleischauer, M.; Ludwig, M.; Hoffmann, M. A.; Rousu, J.; Dorrestein, P. C.; Böcker, S. Classes for the masses: Systematic classification of unknowns using fragmentation spectra. *bioRxiv* **2020**, 2020.04.17.046672. DOI: 10.1101/2020.04.17.046672

14. Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S., Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceut.* **2017**, *14* (12), 4462-4475.

15. Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics*. **2016**, *8*, DOI: 10.1186/s13321-016-0174-y

16. Welker, M.; von Döhren, H. Cyanobacterial peptides - nature's own combinatorial biosynthesis. *FEMS Microbiol.* **2006**, *30* (4), 530-563.

17. Faltermann, S.; Zucchi, S.; Kohler, E.; Blom, J. F.; Pernthaler, J.; Fent, K. Molecular effects of the cyanobacterial toxin cyanopeptolin (CPI 020) occurring in algal blooms: Global transcriptome analysis in zebrafish embryos. *Aquat. Toxicol.* **2014**, *149*, 33-39.

18. Morrison, K. C.; Hergenrother, P. J. Natural products as starting points for the synthesis of complex and diverse compounds. *Nat. Prod. Rep.* **2014**, *31* (1), 6-14.

19. Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *P. Natl. Acad. Sci. U.S.A.* **2005**, *102* (48), 17272-17277.

20. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G., Retrospective analysis of natural products provides insights for future discovery trends. *P. Natl. Acad. Sci. U.S.A.* **2017**, *114* (22), 5601-5606.

21. Cao, Y. Q.; Jiang, T.; Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **2008**, *24* (13), 366-374.

22. Kumar, A.; Bi, L.; Kim, J.; Feng, D. D. Chapter Five - Machine learning in medical imaging. In *Biomedical Information Technology*, 2nd ed; Feng, D. D.; Ed. Academic Press: **2020**; pp 167-196.

23. Roy, A.; Saraf, S. Limonoids: Overview of significant bioactive triterpenes distributed

in plants kingdom. *Biol. Pharm. Bull.* **2006**, *29* (2), 191-201.

24. Zhang, Y. Y.; Xu, H. Recent progress in the chemistry and biology of limonoids. *Rsc. Adv.* **2017**, *7* (56), 35191-35220.

25. Tian, X.; Li, H.; An, F.; Li, R.; Zhou, M.; Yang, M.; Kong, L.; Luo, J. New Structurally Diverse Limonoids from the Seeds of *Khaya senegalensis*. *Planta medica* **2017**, *83* (03/04), 341-350.

26. Reher, R.; Kim, H. W.; Zhang, C.; Mao, H. H.; Wang, M. X.; Nothias, L. F.; Caraballo-Rodriguez, A. M.; Glukhov, E.; Teke, B.; Leao, T., et al. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J. Am. Chem. Soc.* **2020**, *142* (9), 4114-4120.

27. Merwin, N. J.; Mousa, W. K.; Dejong, C. A.; Skinnider, M. A.; Cannon, M. J.; Li, H. X.; Dial, K.; Gunabalasingam, M.; Johnston, C.; Magarvey, N. A. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *P. Natl. Acad. Sci. U.S.A.* **2020**, *117* (1), 371-380.

28. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y. W.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep.* **2017**, *7*, DOI: 10.1038/s41598-017-13923-x

29. Zheng, S. J.; Yan, X.; Gu, Q.; Yang, Y. D.; Du, Y. F.; Lu, Y. T.; Xu, J. QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J. Cheminformatics* **2019**, *11*, DOI: 10.1186/s13321-019-0328-9

30. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436-444.

31. de Matos, P.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Turner, S.; Steinbeck, C. ChEBI: a chemistry ontology and database. *J. Cheminformatics* **2010**, *2*, 6. DOI: 10.1186/1758-2946-2-S1-P6

32. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L. Y.; He, J. E.; He, S. Q.; Shoemaker, B. A.; Wang, J. Y.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic. Acids. Res.* **2016**, *44* (1), 1202-1213.

33. Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123-1124.

34. Gu, J. Y.; Gui, Y. S.; Chen, L. R.; Yuan, G.; Lu, H. Z.; Xu, X. J. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *Plos One* **2013**, *8* (4). DOI: 10.1371/journal.pone.0062839

35. Dewick, P. M. *Medicinal Natural Products: A Biosynthetic Approach*, 3rd ed.; John Wiley & Sons: Chichester, **2009**.
36. Fattorusso, E.; Gerwick, W. H.; Taglialatela-Scafati, O. *Handbook of Marine Natural Products*. Springer: Dordrecht **2012**.
37. Kinghorn, A. D.; Falk, H.; Gibbons, S.; Kobayashi, J. i. *Progress in the Chemistry of Organic Natural Products 106*. Springer International Publishing: Cham, **2017**.
38. Averagesch, N. J. H.; Krömer, J. O. Metabolic Engineering of the Shikimate Pathway for Production of Aromatics and Derived Compounds—Present and Future Strain Construction Strategies. *Front. Bioeng. Biotechnol.* **2018**, *6* (32), DOI: 10.3389/fbioe.2018.00032
39. Downer, E. J. Anti-inflammatory Potential of Terpenes Present in *Cannabis sativa* L. *Acs. Chem. Neurosci.* **2020**, *11* (5), 659-662.
40. Kittakoop, P.; Mahidol, C.; Ruchirawat, S. Alkaloids as important scaffolds in therapeutic drugs for the treatments of cancer, tuberculosis, and smoking cessation. *Curr. Top. Med. Chem.* **2014**, *14* (2), 239-252.
41. Rahman, S. U.; Ismail, M.; Khurram, M.; Ullah, I.; Rabbi, F.; Iriti, M. Bioactive Steroids and Saponins of the Genus *Trillium*. *Molecules* **2017**, *22* (12), 10.3390/molecules22122156
42. Kytidou, K.; Artola, M.; Overkleeft, H. S.; Aerts, J. Plant Glycosides and Glycosidases: A Treasure-Trove for Therapeutics. *Front. Plant. Sci.* **2020**, *11*, 357. DOI: 10.3389/fpls.2020.00357
43. Bhat, K. L.; Joullie, M. M. Cyclopeptide Alkaloids. *J. Chem. Educ.* **1987**, *64* (1), 21-27.
44. Landrum, G. RDKit: Open-source cheminformatics from machine learning to chemical registration. *Abstr. Pap. Am. Chem. Soc.* **2019**, 258.
45. Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (3), 1253-1268.
46. Goodin, S. Ixabepilone: a novel microtubule-stabilizing agent for the treatment of metastatic breast cancer. *Am. J. Health. Syst. Pharm.* **2008**, *65* (21), 2017-2026.
47. Clark, D. E. Virtual Screening: Is Bigger Always Better? Or Can Small Be Beautiful? *J. Chem. Inf. Model.* **2020**, DOI: doi.org/10.1021/acs.jcim.0c00101
48. Monaghan, P.; Fardis, M.; Revill, W. P.; Bell, A. Antimalarial Effects of Macrolactones Related to FK520 (Ascomycin) Are Independent of the Immunosuppressive Properties of the

Compounds. *J. Infect.* **2005**, *191* (8), 1342-1349.

49. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z., et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688-702.

50. Sestili, P.; Stocchi, V. Repositioning Chromones for Early Anti-inflammatory Treatment of COVID-19. *Front Pharmacol* **2020**, *11*, 854-854.

51. Su, H.; Yao, S.; Zhao, W.; Li, M.; Liu, J.; Shang, W.; Xie, H.; Ke, C.; Hu, H.; Gao, M.; Yu, K.; Liu, H., et al. Anti-SARS-CoV-2 activities in vitro of Shuanghuanglian preparations and bioactive ingredients. *Acta. Pharmacologica. Sinica.* **2020**.

52. Garg, N.; Luzzatto-Knaan, T.; Melnik, A. V.; Caraballo-Rodríguez, A. M.; Floros, D. J.; Petras, D.; Gregor, R.; Dorrestein, P. C.; Phelan, V. V. Natural products as mediators of disease. *Nat. Prod. Rep.* **2017**, *34* (2), 194-219.

Graphical abstract