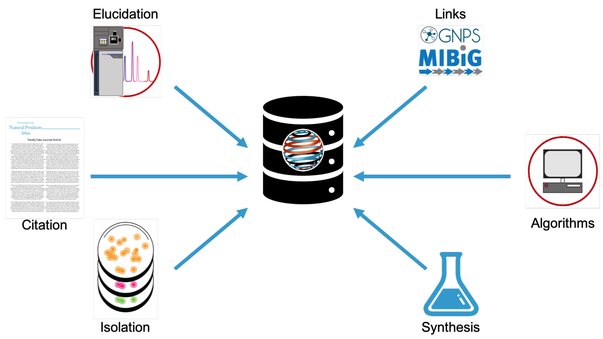
The Natural Products Atlas:
An Open Access Platform for Natural Products Discovery

Jeffrey A. van Santen, Roger G. Linington*
Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6

Scope

The Natural Products Atlas (NP Atlas) aims to be the complete, community-driven, free for academic use database of all known microbial natural products.

The NP Atlas requires deep community involvement and support.



Compounds are the central pillar of the NP Atlas. Specifically, the database includes data on all non-primary metabolite,[1] small molecules discovered from bacteria and fungi published in the peer-reviewed primary scientific literature. This encompasses bacterial, fungal and cyanobacterial compounds, but does not include compounds from plants, invertebrates or other higher organisms unless these compounds have also been explicitly identified from a microbial source. Compounds from lichens and mushrooms and other higher fungi are included. Compounds from marine macro algae and diatoms are excluded. We have collected and manually curated data on the original isolation and structural elucidation of each compound. We also aim to have the have the most up-to-date structure of each compound. The NP Atlas also contains some citation information on total syntheses of compounds and instances of structural revisions. The database is versioned and a download dump is freely available.

Today, the NP Atlas has an estimated 80% coverage of microbial natural products with original isolation names and producing organisms.

We also aim to be a central repository for natural product research. To facilitate and ease research, we have created and continue to integrate links with other valuable databases and tools, such as MIBiG and GNPS.[2,3]

The NP Atlas also features online forms for submitting data and corrections, as well as a new online data curation platform for seamless collaboration with worldwide volunteers.

Future Developments

The long-term vision of the NP Atlas is to contain the following essential data:

- All microbial natural products published in the primary literature and patents
- A record of all published names and synonyms for each compound
- All instances of total synthesis of each compound
- All instances of structural revision of each compound
- Physicochemical data including IR and UV spectra, along with direct links to NMR and MS databases
- Complete taxonomic descriptions, including higher designations (phylum, order, etc.), and a full list of producing organisms for each compound
- Incorporate information on bio-activity

In addition, we have a variety of plans to improve the number of use cases for the Atlas, as well as to ease use.

- Increase the number and variety of *Discover* dashboards
- Build a full RESTful API to access and deposit data
- With appropriate resources, the NP Atlas is well poised to extend it scope to marine invertebrates and ultimately plants

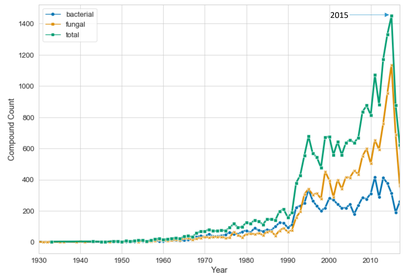
References

(1) As defined by KEGG: Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27-30. **(2)** Wang, M., et al. *Nature Biotechnol.* **2016**, *34.8*, 828-837. **(3)** Madema, MH.; et al. *Nature Chem. Biol.* **2015**, *11*, 625-631. **(4)** RDKit: Open-source cheminformatics; http://www.rdkit.org **(5)** ORCID: https://orcid.org **(6)** Kim, S., et. Al. *Nucleic Acids Res.* **2019,** *47,* D1102-1109. **(7)** Sayers, E.W., et al. *Nucleic Acids Res.* **2019**, *47,* D23-28. **(8)** Other data has been gather from a variety of other legacy databases. **(9)** Jiang, C., et al. *J. Chem. Inf. Model.* **2016**, *56,* 1132-1138. **(10)** Hähnke, V.D., et al. *J. Cheminform.* **2018**, *10,* 36. **(11)** Python 3: https://www.python.org/ **(12)** Flask: https://palletsprojects.com/p/flask/ **(13)** MySQL: https://www.mysql.com/ **(14)** Jinja2: http://jinja.pocoo.org/

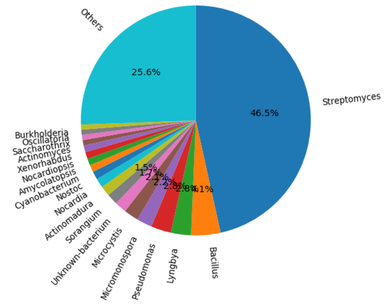
Database Statistics

Current # of Compounds: _____

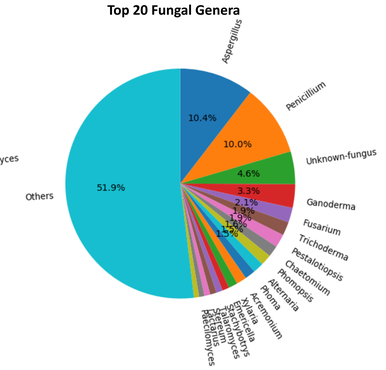
 # Compounds per Year



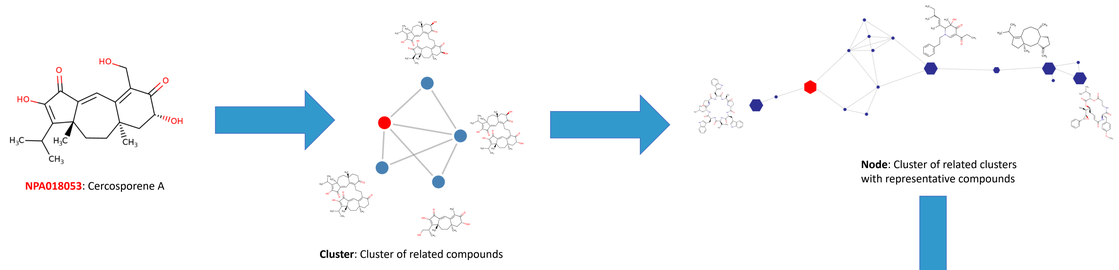
Top 20 Bacterial Genera



Top 20 Fungal Genera



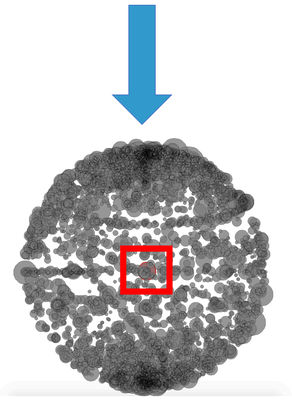
Multitiered Viewpoints of Chemical Space



NPA018053: Cercosporene A

Cluster: Cluster of related compounds

Node: Cluster of related clusters with representative compounds



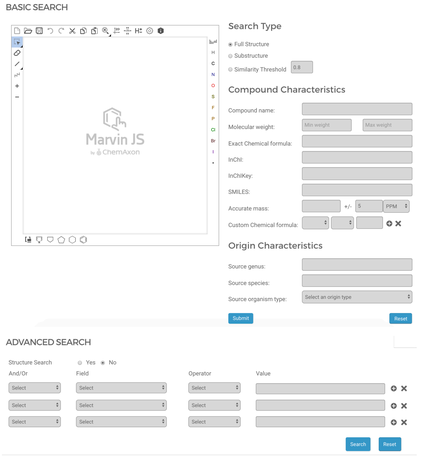
Global: Nodes displayed on interactive spherical plot

Interactive tools on the NP Atlas website allow for the exploration of known natural product chemical space. The above figures demonstrate tracking Cercosporene A (NPA018053) through each of the various tiers of organization, as highlighted in red.

Compounds are given a static NPAID when deposited into the database. Before the release of a new version of the database, a series of algorithms are run to create three additional viewpoints to examine related compounds. Clusters are a useful way to find compounds that are closely related. Clusters are created by using Morgan fingerprinting (radius = 2) and Dice similarity scoring in the RDKit,[4] with a similarity cutoff of 0.75. Nodes are groups of clusters that share more general structural features. Nodes are calculated by finding the most interconnected compound in each cluster, and using these compounds as representatives for each cluster to calculate structure similarities. This calculation differs from cluster similarities in both the fingerprinting method (Atom pairs) and the similarity scoring method (Dice, 0.7 cutoff). The global view presents all of the nodes in the NP Atlas, arrayed as a spherical plot. The distribution of nodes in this spherical plot is defined based on their C:H, C:O and C:N ratios. C:H defines the polar angle. C:O defines the azimuth value and the C:N ratio sets the radius (distance from the origin).

Search results can be displayed over the global view, making it an interactive playground for exploring known chemical space or relating chemical scaffolds.

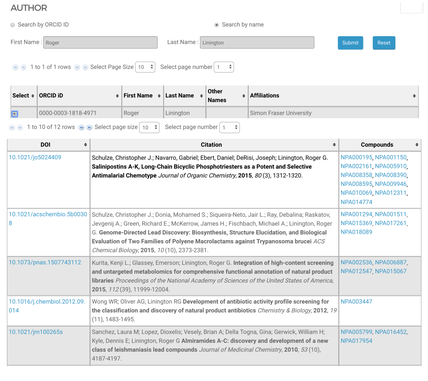
The NP Atlas is Highly Searchable and Explorable



At npatlas.org, there are a variety of methods for searching and exploring the data. *Basic Search* is the primary page for finding compounds by name, chemical properties, origin information, structures, and sub-substructures.

The *Advanced Search* page on the other hand allows for more advanced queries. For example, you can search for a specific article be title or identifier (DOI/PMID). Or you can perform detailed Boolean searches, such as, compounds with molecular weight greater than three hundred with more than one chlorine atom and not containing nitrogen and published between 1990 and 2001.

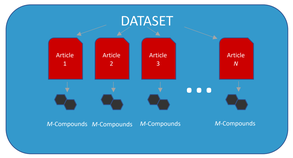
There are a variety of other dashboards under the *Discover* menu which allow for exploring the database. One such page, the *Discover Author* page, allows for searching by ORCID ID,[5] or author name if you don't know the ID.



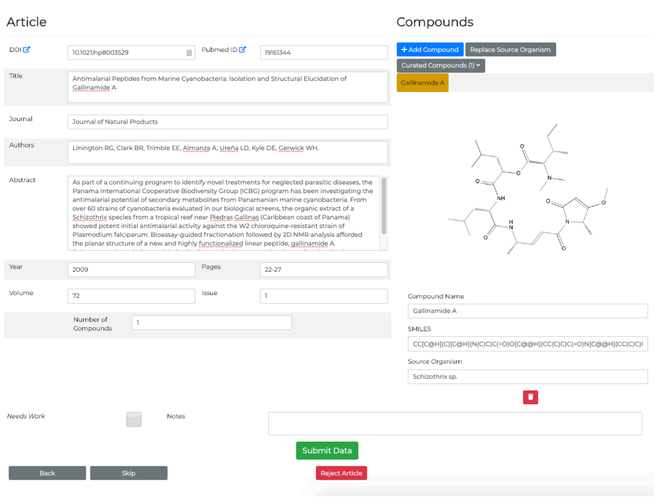
Data Curation Model and Platform

Initially, the NP Atlas was built by hand curating articles from a top 50 priority list of journals, dated from 2015 back to about 1995. These initial results were then used to build a linear classifier machine learning model capable of determining whether an article relates to natural product isolation. This model is limited by the free availability of only title and abstract strings. The initial model is promiscuous: it has a very low false negative rate, but a rather high false positive rate. Therefore, articles are currently only considered if the title or abstract contain bacterial or fungal genera terms in, excluding pathogenic terms.

The current data workflow is as follows: Data is gather from a variety of sources including PubChem and PubMed,[6-8] and then compiled into a standardized format. The data is then filtered using our machine learning model described above. The data is then hand curated by our team, followed by an additional quality control check, done by our in-house "checker" algorithm. Finally, data is inserted into a development version of the NP Atlas, until the next release cycle.



As each entry of the NP Atlas is carefully hand curated by our team, we required software which enables us to be consistent, while easing the process as much as possible. After several prototypes, we have now launched a web-based curation platform at npatlas-curate.chem.sfu.ca. This database backed software enables us to easily collaborate with international colleagues. The data is organized in a hierarchical manner, with each dataset containing *N*-articles, which each contain *M*-compounds. This system allows us to easily organize data which needs reviewing into segmented pieces, and allows us to easily assign datasets to our volunteers.



The online curation platform features an article-by-article design, with a detailed form for all citation information, including DOIs and PubMed IDs, that link-out to the article for rapid validation. Compound information is displayed in a tabbed viewer. Compound structures are collected using isomeric SMILES strings, and are rendered and displayed on the web using the Kekule.js JavaScript library.[9] Compound structures are also standardized using the PubChem Standardization public resource.[10] Compound names and source organisms are regularized in the post-curation "checker" algorithm. The website is built using a the Flask framework for Python, and has a MySQL database backend. The frontend is rendered using a combination of the Jinja2 template engine that generates HTML, and custom built CSS, and JavaScript code.[11-14]

If you'd like to volunteer as a curator, please email us at: volunteer@npatlas.org

Acknowledgements

The NP Atlas could not exist with out the support of our past, present, of hopefully future curators: Victor Aniebok, Marcy Balunas, Derek Bunsko, Fausto Carnevale-Neto, Laia Castano, Steve Chang, Trevor Clark, Jessica Cleary, David Delgadillo, Katherine Duncan, Joseph Egan, Claire Fergusson, Jackie Fries, Melissa Galey, Jake Haeckl, Alex Hua, Alison Hughes, Dasha Iskakova, Grégoire Jacob, Aswad Khadilkar, Kenji Kurita, Jeongho Lee, Sanghoon Lee, Nicole LeGrow, Roger Linington, Dennis Liu, Jocelyn Macho, Catherine McCaughey, Jason McFarlane, Ram P Neupane, Timothy O'Donnell, Laura Sanchez, Anam Shaikh, Amrit Leen Singh, Sylvia Soldatou, Tuan Anh Tran, Mercia Valentine, Jeffrey van Santen, Duy Vo, Darryl Wilson, Katherine Zink