The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery

*Jeffrey A. van Santen, Grégoire Jacob, Amrit Leen Singh, Victor Aniebok, Marcy J. Balunas, Derek Bunsko, Fausto Carnevale Neto, Laia Castaño-Espriu, Chen Chang, Trevor N. Clark, Jessica L. Cleary Little, David A. Delgadillo, Pieter C. Dorrestein, Katherine R. Duncan, Joseph M. Egan, Melissa M. Galey, F.P. Jake Haeckl, Alex Hua, Alison H. Hughes, Dasha Iskakova, Aswad Khadilkar, Jung-Ho Lee, Sanghoon Lee, Nicole LeGrow, Dennis Y. Liu, Jocelyn M. Macho, Catherine S. McCaughey, Marnix H. Medema, Ram P. Neupane, Timothy J. O'Donnell, Jasmine S. Paula, Laura M. Sanchez, Anam F. Shaikh, Sylvia Soldatou, Barbara R. Terlouw, Tuan Anh Tran, Mercia Valentine, Justin J. J. van der Hooft, Duy A. Vo, Mingxun Wang, Darryl Wilson, Katherine E. Zink, Roger G. Linington**

Supporting Information

Table of Contents

SI: The Natural Products Atlas Curation Platform 2

SII: Quality Control Platform 6

References 10

List of Figures

Figure S1: The Natural Products Atlas data curation simplified workflow. 2
Figure S2: Hierarchical data organization used in the Natural Products Atlas curation platform. 3
Figure S3: Screenshot of the Natural Products Atlas online curation platform. 4
Figure S4: Checker algorithm flowchart. Red indicates rejection, green indicates success, yellow indicates an issue which has arisen, and orange indicates data associated with the manual review platform. 8
Figure S5: Example quality control issue resolution web-form. In this web-form, a compound was flagged as a duplicate with respect to the current version of the Natural Products Atlas. This is an exact duplicate, and therefore can be rejected from future consideration. 9

SI: The Natural Products Atlas Curation Platform

Initially, the Natural Products Atlas was built by hand curating articles from a selection of the top 30 journals known to contain articles reporting novel natural product discovery, as described in more detail in the primary manuscript. The results of these efforts were ~30,000 articles annotated as relating to natural product isolation, or not, with the overall set containing 12,924 compounds. These initial results were used to train a support vector machine, linear classifier machine learning model capable of determining whether or not an article relates to natural product isolation. This model is limited by the free availability of only title and abstract strings. The model is promiscuous; however, it serves its purpose well as it has a very low false negative rate, but a rather high false positive rate. To combat this, articles are currently only considered if the title or abstract contain bacterial or fungal genera terms, excluding pathogenic genera.

The current data workflow, simplified in *Figure S1*, is as follows: data are gathered from a variety of sources including PubChem and PubMed,[1,2] and then compiled into a standardized format. The data are then filtered using the machine learning model described above. The data are then hand curated by our team, followed by an additional quality control checking stage, performed by our in-house "checker" algorithm which is described below in section *SII: Quality Control Platform*. Finally, data are inserted into a development version of the Natural Products Atlas, and deployed to the production version at the next quarterly release cycle.

Figure S1*: The Natural Products Atlas data curation simplified workflow.*

Each entry in the Natural Products Atlas is carefully hand curated by our team. To tackle this extensive task, we required software which would ensure consistency, while facilitating and streamlining the curation process. After several prototypes, we have now launched a web-based

curation platform at npatlas-curate.chem.sfu.ca. The data are organized in a hierarchical structure, with each dataset containing *N*-articles and each article containing *M*-compounds, as show in *Figure S2*. This system allows data to be easily organized into segmented pieces and facilitates the assignment of datasets to curators. Data are managed by an administrator, who inserts data into the curation platform, and has full control over the quality control platform.

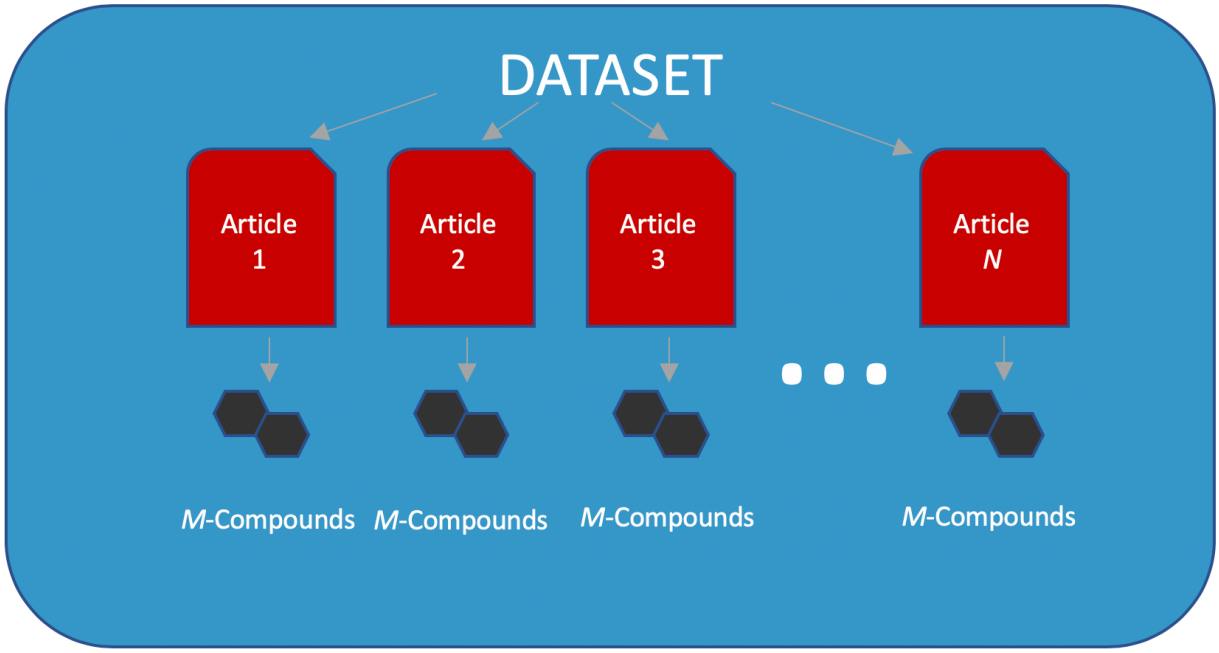

Figure S2: *Hierarchical data organization used in the Natural Products Atlas curation platform.*

The online curation platform features an article-by-article design, with a detailed form for each candidate article, containing all citation information, including titles, abstracts, authors, publication year, volume, issue, pages, journal, and DOIs and PubMed IDs that link-out to the article for rapid validation. A sample screenshot of the curation software is shown in *Figure S3*. Compound information is displayed in a data-rich, tabbed viewer. Compound structures are collected using isomeric SMILES strings, and are rendered on the web using the Kekule.js JavaScript library.[3] Prior to curation, compound structures are standardized using the PubChem Standardization public resource.[4]

The website backend is built using the Flask framework for Python and uses a MySQL database. The frontend is rendered using a combination of the Jinja2 template engine, which generates HTML, and custom-built CSS, and JavaScript code.

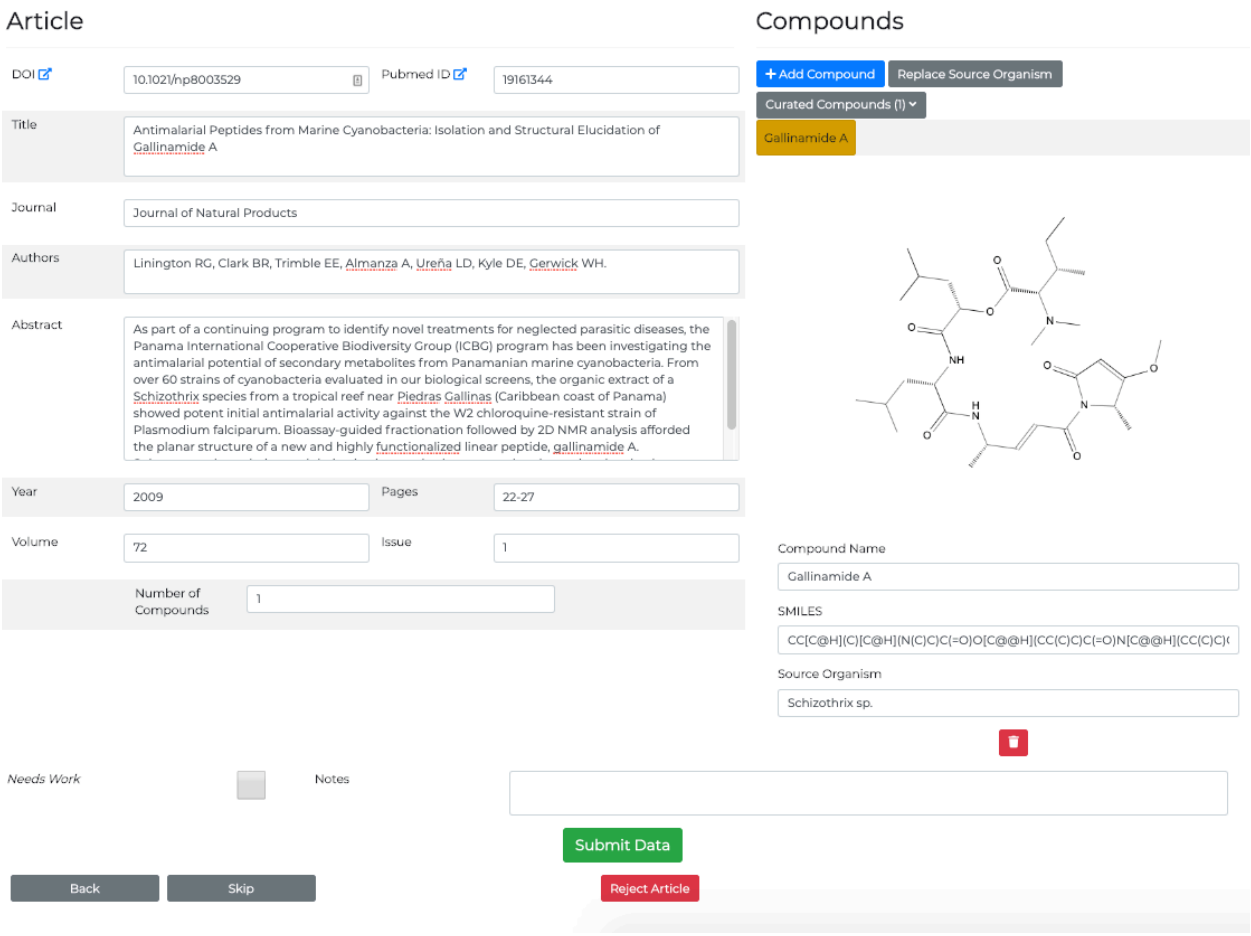

Figure S3: *Screenshot of the Natural Products Atlas online curation platform.*

Volunteer curators use the curation platform in the following manner. The landing page for their account displays the datasets currently assigned to them, as well as the degree of completion of each dataset. Once a dataset has been selected for curation, the next uncurated article in the set is displayed. Curators review the abstract to ensure that the paper describes novel microbial natural products discovery. If it does not, the article is rejected, and the next article is displayed. If it does report novel natural product discovery the curator reviews the citation data for accuracy/ completeness, and either reviews the producing organism (if pre-populated) or enters these data. Finally, the curator

examines the structure panel, and ensures that the correct compounds are listed, and that structures exist for each one. If not, there are tools for adding and removing compounds, as well as correcting both compound names and structures. Linkouts to both the article (via the DOI), and PubMed are available to facilitate data review. Once all data are correct, the curator clicks 'Submit Data' and the next article in the dataset is displayed. Occasionally articles are either confusing, or outside the technical expertise of the curator. In these cases, curators may check the 'Needs Work' check box, and add a note in the adjacent text box describing the issue.

SII: Quality Control Platform

Quality control and data consistency are important tenets of the Natural Products Atlas. As such, additional steps are taken post curation to ensure the highest possible quality of data, as well as to avoid duplication within the database. Firstly, to ensure consistency, all structures which pass curation are once again standardized using the PubChem Standardization service.[4] Next, the quality control algorithm, which we call the "checker", is applied to all articles and compounds which have passed curation. This algorithm is outlined in *Figure S4*. As described above, each curation dataset is structured hierarchically. This is leveraged in the quality control platform by checking each completed dataset. Once a dataset has been pulled into the checker by an administrator, a curator can no longer review or alter the dataset.

A detailed description of the checker algorithm is as follows: For each of *N*-articles in a dataset, the article is queried for having passed curation. Non-passed articles can exist because an article was flagged as rejected for not being about Natural Products isolation, or as needing additional review. The citation information is validated and regularized. This involves verifying that journal titles are listed in our database, validating year strings and DOIs using regular expressions, and checking that all other citation information is not corrupted. Each article is then queried against the Natural Products Atlas database to avoid duplication. If a duplicate is found, the data are merged and checking continues.

Next, for each of *M*-compounds in an article, structural information is loaded using the RDKit library for Python.[5] If any fragments or salts are detected, they are stripped and the compound is flagged for manual review. Next, compound name formatting is regularized. Compounds with no reported name are all called "Not named", names are capitalized if appropriate, suffices such as "acid", "ether", etc., are de-capitalized, and suffices such as "A", "B1", etc., are capitalized. The next steps are crucial in keeping duplicate compounds out of the Natural Products Atlas. Compounds and articles may sometimes be re-curated using the curation/checker platforms depending on the original search terms

used; therefore, these compounds are checked for any structural changes from the current version of the database. If any are detected, a manual review is prompted. If a manual review has already been performed, the previous step is omitted. A fuzzy search of the database is also performed by looking for all instances of a compound's name matching the database (excluding "Not named"), searching for complete structural overlap including stereochemistry, using InChIKeys,[6] and searching for flat structure matches, also using InChIKeys. If any of these matches are made, manual review is prompted.

Next, compounds are verified as 'natural product like' according to our criteria. Compounds with molecular weights greater than 3000 Dalton are rejected. Compounds which have previously been identified as retracted natural products are rejected. Fluorinated compounds are flagged for manual review. If any flags have been raised at any point during the checker algorithm, these problems are manually inspected by an administrator. This review is performed using web-forms tailored to each specific problem encountered. An example web-form is demonstrated in *Figure S5*. Articles are re-checked after issue resolution to make sure no new problems were introduced by the administrator. At this point, the dataset has passed quality control and is inserted into the development database in preparation for the next quarterly review cycle.

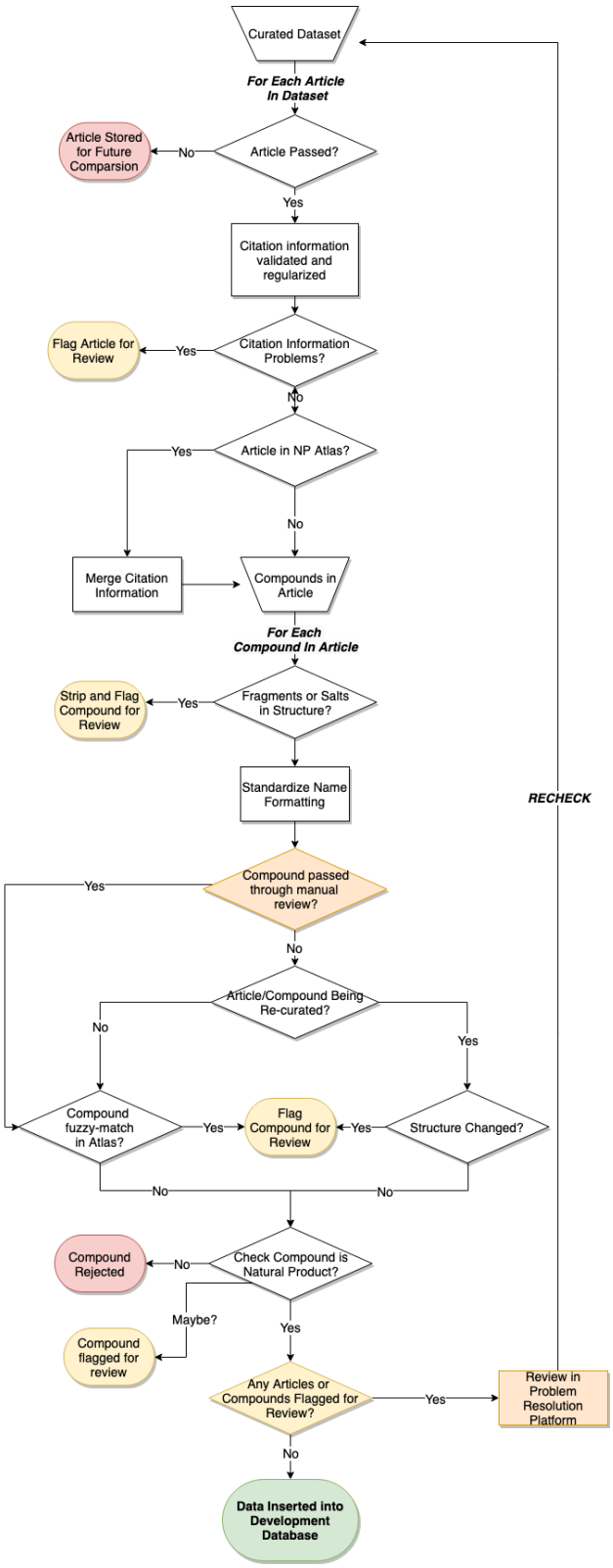

Figure S4: *Checker algorithm flowchart. Red indicates rejection, green indicates success, yellow indicates an issue which has arisen, and orange indicates data associated with the manual review platform.*

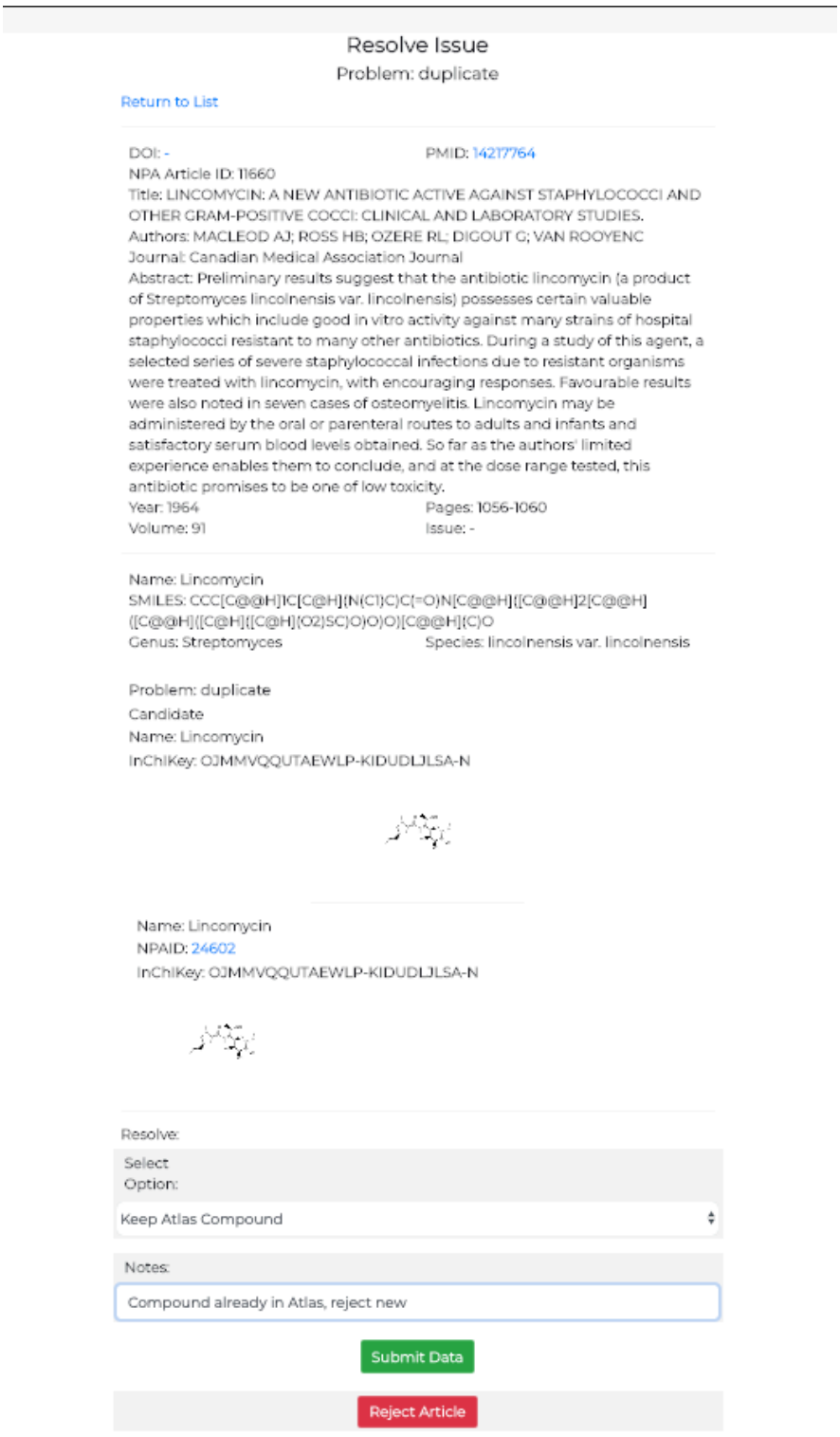

Figure S5: *Example quality control issue resolution web-form. In this web-form, a compound was flagged as a duplicate with respect to the current version of the Natural Products Atlas. This is an exact duplicate, and therefore can be rejected from future consideration.*

References

(1) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. Nucleic Acids Res. 2019, 47 (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(2) Sayers, E. W.; Agarwala, R.; Bolton, E. E.; Brister, J. R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019, 47 (D1), D23–D28. https://doi.org/10.1093/nar/gky1069.

(3) Jiang, C.; Jin, X.; Dong, Y.; Chen, M. Kekule.Js: An Open Source JavaScript Chemoinformatics Toolkit. J. Chem. Inf. Model. 2016, 56 (6), 1132–1138. https://doi.org/10.1021/acs.jcim.6b00167.

(4) Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem Chemical Structure Standardization. J. Cheminformatics 2018, 10 (1), 36. https://doi.org/10.1186/s13321-018-0293-8.

(5) RDKit: Open-Source Cheminformatics; http://www.rdkit.org.

(6) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. J. Cheminformatics 2013, 5, 7. https://doi.org/10.1186/1758-2946-5-7.