The Journal of Open Source Software

NPLinker 2: a modular and customizable framework for paired omics analyses

Cunliang Geng [1¶], **Giulia Crocioni** [1], **Helge Hecht** [2], **Arjan Draisma** [3], **Annette Lien** [3], **Laura Rosina Torres Ortega** [3], **Dora Ferreira** [4], **Pablo Lopez-Tarifa** [1], **Katherine R. Duncan** [5], **Marnix H. Medema** [3], and **Justin J. J. van der Hooft** [3,6¶]

1 Netherlands eScience Center, Netherlands **2** RECETOX, Faculty of Science, Masaryk University, Kotlářská 2, 60200, Brno, Czech Republic **3** Bioinformatics Group, Wageningen University & Research, Netherlands **4** Naicons Srl, Milan, Italy **5** Newcastle University, Biosciences Institute, Newcastle upon Tyne, UK **6** Department of Biochemistry, University of Johannesburg, 2006 Johannesburg, South Africa

¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Evan Spotte-Smith

Reviewers:

- @bgyori
- @apraga

Submitted: 11 August 2025
Published: unpublished

Summary

Natural product discovery increasingly relies on the integration of multi-omics data to explore and prioritize biochemical diversity. To advance these efforts, we present NPLinker 2, a redesigned Python framework to do paired omics analyses by prioritizing genomics-metabolomics links. It provides a modular workflow that allows defining custom modules for data preparation, data loading and scoring methods. In addition, NPLinker 2 includes a web application for the interactive analysis and visualisation of promising links.

Statement of need

Omics datasets have become a key resource for natural products discovery, enabling the systematic exploration of specialized metabolites, the refinement of knowledge of known natural products, and the identification of novel bioactive compounds or metabolic enzymes. Paired omics analyses combine complementary genomics (e.g., biosynthetic gene clusters (BGCs)) and metabolomics (e.g., mass spectra) datasets to elucidate gene-metabolite relationships, accelerating the discovery process (Goering et al., 2016; Hooft et al., 2020; Leão et al., 2022). However, omics data structures, preproccessing pipelines, resources, and annotation tools are constantly being improved. For example, newer releases of MIBiG contain more validated BGCs and new annotation fields (Zdouc et al., 2025), while mass spectral libraries are growing in size and information as well (Wang et al., 2016). Besides, newer versions of omics clustering tools have different output file formats. Together with the constant expansion of available experimental datasets, this puts a strain on downstream frameworks that integrate the data and results. Hence, natural products discovery would benefit from up-to-date and user-friendly software packages that parse processed omics data and connect it with algorithms returning ranked, queryable gene cluster - mass spectra links to prioritize links to further investigate manually. Here, we redesigned NPLinker to provide such an integrative omics tool that guides both users and developers in paired omics mining with its modular setup. For example, recent developments in omics processing, annotation tools, and ranking metrics could be added to the framework (Louwen, Medema, et al., 2023; Louwen, Kautsar, et al., 2023). Moreover, several of such linking scores could then be used together with the currently implemented strain correlation score to further improve ranking results.

The Journal of Open Source Software

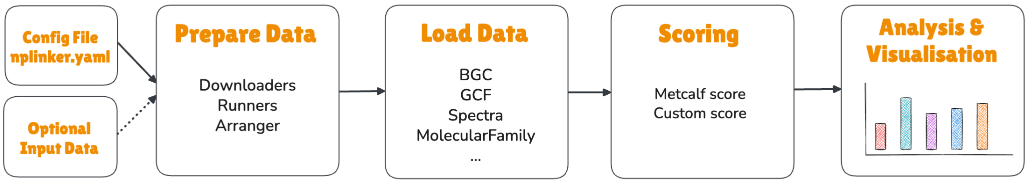

Figure 1: The NPLinker 2 framework. The current pipeline consists of five main components: 1. Initiating an analysis with an input block that includes configuration file and optional input data; 2. Preparing dataset by automatically downloading or generating data; 3. Loading and parsing data from data files; 4. Scoring and linking data; 5. Creating an output for analysis and visualization of results.

Features of NPLinker 2

NPLinker 2 is redesigned based on NPLinker version 1.x (Eldjárn et al., 2021) to provide a more flexible, modular and extensible framework for linking BGCs to mass spectra. The pipeline is shown in the Figure 1, and the key features are highlighted below.

Installation ease

NPLinker 2 is distributed as a Python package, but it relies on several third-party tools and databases that are not available via PyPi, which can make the installation more complex. To simplify the setup process, NPLinker 2 includes an installation script that automatically installs the required non-PyPi dependencies and databases.

To install NPLinker 2 and its dependencies, users can run the following commands:

```
# Install the NPLinker package
pip install --pre nplinker

# Install non-PyPi dependencies and required databases
install-nplinker-deps
```

Configurable

NPLinker 2 can be easily configured using the file `nplinker.yaml`, which is required to customise the pipeline according to users' needs, e.g., by selecting the run mode, choosing scoring methods. A friendly template is available on the doc website to help users create and fill in their configuration file from scratch.

Local mode and PODP mode

NPLinker 2 supports both local and remote data sources through two operational modes: **local mode** and **PODP mode**. In local mode, users provide their local data files, e.g., AntiSMASH (Blin et al., 2025) output files and mass spectral data (Wang et al., 2016) supported by matchms (Huber et al., 2020), as input. In contrast, the Paired Omics Data Platform (PODP) (Schorn et al., 2021) mode requires no input data files from the user, as the pipeline automatically downloads necessary data files from the PODP (Paired Omics Data Platform) server using the PODP ID specified in the `nplinker.yaml` file. This dual-mode support enables private and public data analysis using the same pipeline.

Modular and extensible

Modularity and extensibility are key features of NPLinker 2, which provides a set of interfaces and data models that users can extend.

“Prepare Data” component: The core class of this component, `DatasetArranger`, orchestrates various downloaders and runners to automatically download and generate the required data files. These files are then stored in the local working directory specified in the `nplinker.yaml` configuration file. Users can extend this component by adding new downloaders or runners, e.g., to download BGC data from a new source or generate data files using a different method or tool.

“Load Data” component: The `DatasetLoader` class manages data loaders responsible for loading and parsing genomics, metabolomics, and strain data files. Users can add new data loaders to support additional sources or formats. For example, to load BGC data from a new source, one can define a Python class `NewBGCLoader` that inherits from the `BGCLoaderBase` interface and implements the `get_files` and `get_bgcs` methods, then register it within the `DatasetLoader` class.

“Scoring” component: This component handles the linking of data and the scoring of those links. A undirected graph is used to store the linked data, with nodes corresponding to genomics or metabolomics data items and edges representing the links between them with scoring values, as illustrated in Figure 2. The `ScoringBase` interface is provided to allow the implementation of custom scoring methods.

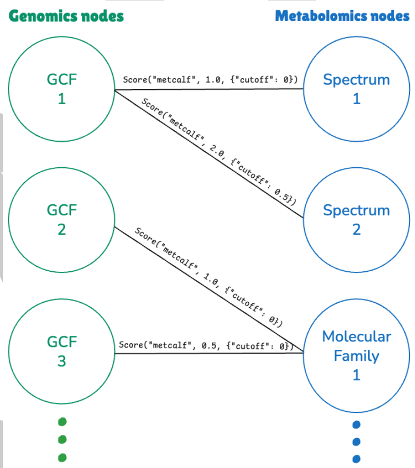

Figure 2: Graph representation of linkings.

New documentation website

A dedicated documentation website is available to help users and developers understand how to use and extend NPLinker 2. It includes tutorials, conceptual overviews, pipeline diagrams, and an API reference. The documentation is available at https://nplinker.github.io/nplinker.

New unit tests and integration tests

Unit and integration tests are included in NPLinker 2 to ensure the codebase and the overall pipeline is correct. The tests can be run in parallel to speed up the testing process.

Forced static typing

NPLinker 2 is developed with forced static typing, which means that all functions and methods have type hints to specify the input and output types. This helps developers to understand the code better and catch type errors early when dealing with complex genomic and metabolomic data and the processed and annotated data derived thereof.

User-friendly webapp

The NPLinker web application (webapp) is an interactive dashboard built with Plotly Dash, designed to make NPLinker's linking results accessible through a user-friendly web interface. A publicly hosted demo allows immediate testing: users can load a sample dataset with a single click and start exploring. To enable full functionality with larger datasets, the webapp can be installed locally or run via Docker (using the nplinker-webapp image). Notably, the webapp focuses on visualization and post-analysis; link scoring between genomic and metabolomic entries is performed beforehand by the NPLinker backend.

The linking is currently provided starting from both omics views: from Gene Cluster Families (GCFs) clustered by BiG-SCAPE (Navarro-Muñoz et al., 2020) to mass spectra or molecular families (MFs) clustered by molecular networking (Wang et al., 2016) or vice vera. Once the data is loaded, the interface provides two complementary views: genomics-to-metabolomics and metabolomics-to-genomics. This dual-tab layout allows users to begin from either data type and inspect associated links in the other domain. Each view presents the input data and predicted links in sortable, filterable tables, with support for multiple filtering criteria (e.g., GCF, MF, spectrum IDs, BGC classes, score thresholds). This enables rapid prioritization of promising BGC–metabolite links. Results can also be exported as Excel files for downstream analysis and record-keeping, allowing smooth integration into existing workflows.

Acknowledgements

This work was supported by the Netherlands eScience Center under grant number NLESC.OEC.2021.002 (M.H. Medema and J.J.J. van der Hooft). A. Lien, L. R. Torres Ortega, D. Ferreira, K.R. Duncan, M.H. Medema and J. J. J. van der Hooft acknowledge the MAGic-MOLFUN Doctoral Training Network that has received funding from the European Union's Horizon Europe programme under the Marie Skłodowska-Curie grant agreement No. 101072485 (M.H. Medema and J.J.J. van der Hooft) and and UKRI EP/X03142X/2 (K.R. Duncan).

Conflict of Interest

JJJvdH is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA. MHM is a member of the scientific advisory boards of Hexagon Bio and Hothouse Therapeutics Ltd. All other authors declare to have no competing interests.

Reference

Blin, K., Shaw, S., Vader, L., Szenei, J., Reitz, Z. L., Augustijn, H. E., Cediel-Becerra, J. D. D., Crécy-Lagard, V. de, Koetsier, R. A., Williams, S. E., Cruz-Morales, P., Wongwas, S., Segurado Luchsinger, A. E., Biermann, F., Korenskaia, A., Zdouc, M. M., Meijer, D., Terlouw, B. R., Hooft, J. J. J. van der, … Weber, T. (2025). antiSMASH 8.0: Extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Research*, *53*, W32–W38. https://doi.org/10.1093/nar/gkaf334

Eldjárn, G. H., Ramsay, A., Hooft, J. J. J. van der, Duncan, K. R., Soldatou, S., Rousu, J., Daly, R., Wandy, J., & Rogers, S. (2021). Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLOS Computational Biology*, *17*, e1008920. https://doi.org/10.1371/journal.pcbi.1008920

Goering, A. W., McClure, R. A., Doroghazi, J. R., Albright, J. C., Haverland, N. A., Zhang, Y., Ju, K.-S., Thomson, R. J., Metcalf, W. W., & Kelleher, N. L. (2016). Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal

Peptide with an Unusual Amino Acid Monomer. *ACS Central Science*, *2*, 99–108. https://doi.org/10.1021/acscentsci.5b00331

Hooft, J. J. J. van der, Mohimani, H., Bauermeister, A., Dorrestein, P. C., Duncan, K. R., & Medema, M. H. (2020). Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews*, *49*, 3297–3314. https://doi.org/10.1039/D0CS00162G

Huber, F., Verhoeven, S., Meijer, C., Spreeuw, H., Castilla, E. M. V., Geng, C., Hooft, J. J. j van der, Rogers, S., Belloum, A., Diblen, F., & Spaaks, J. H. (2020). Matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, *5*, 2411. https://doi.org/10.21105/joss.02411

Leão, T. F., Wang, M., Silva, R. da, Gurevich, A., Bauermeister, A., Gomes, P. W. P., Brejnrod, A., Glukhov, E., Aron, A. T., Louwen, J. J. R., Kim, H. W., Reher, R., Fiore, M. F., Hooft, J. J. J. van der, Gerwick, L., Gerwick, W. H., Bandeira, N., & Dorrestein, P. C. (2022). NPOmix: A machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. *PNAS Nexus*, *1*, pgac257. https://doi.org/10.1093/pnasnexus/pgac257

Louwen, J. J. R., Kautsar, S. A., Burg, S. van der, Medema, M. H., & Hooft, J. J. J. van der. (2023). iPRESTO: Automated discovery of biosynthetic sub-clusters linked to specific natural product substructures. *PLOS Computational Biology*, *19*, e1010462. https://doi.org/10.1371/journal.pcbi.1010462

Louwen, J. J. R., Medema, M. H., & Hooft, J. J. J. van der. (2023). Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. *Microbiome*, *11*, 13. https://doi.org/10.1186/s40168-022-01444-3

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., & Medema, M. H. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology*, *16*, 60–68. https://doi.org/10.1038/s41589-019-0400-9

Schorn, M. A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D. D., Aksenov, A. A., Aleti, G., Moghaddam, J. A., Aron, A. T., Aziz, S., Bauermeister, A., Bauman, K. D., Baunach, M., Beemelmanns, C., Beman, J. M., Berlanga-Clavero, M. V., Blacutt, A. A., Bode, H. B., Boullie, A., … Hooft, J. J. J. van der. (2021). A community resource for paired genomic and metabolomic data mining. *Nature Chemical Biology*, *17*, 363–368. https://doi.org/10.1038/s41589-020-00724-z

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., … Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, *34*, 828–837. https://doi.org/10.1038/nbt.3597

Zdouc, M. M., Blin, K., Louwen, N. L. L., Navarro, J., Loureiro, C., Bader, C. D., Bailey, C. B., Barra, L., Booth, T. J., Bozhüyük, K. A. J., Cediel-Becerra, J. D. D., Charlop-Powers, Z., Chevrette, M. G., Chooi, Y. H., D'Agostino, P. M., Rond, T. de, Del Pup, E., Duncan, K. R., Gu, W., … Medema, M. H. (2025). MIBiG 4.0: Advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research*, *53*, D678–D690. https://doi.org/10.1093/nar/gkae1115