NPLinker Community Meeting

2024-12-04

netherlands eScience center

Agenda

16:00	Announcements and upcoming events
16:05	NPLinker development
16:15	Webapp development
16:25	Q&A
17:00	.

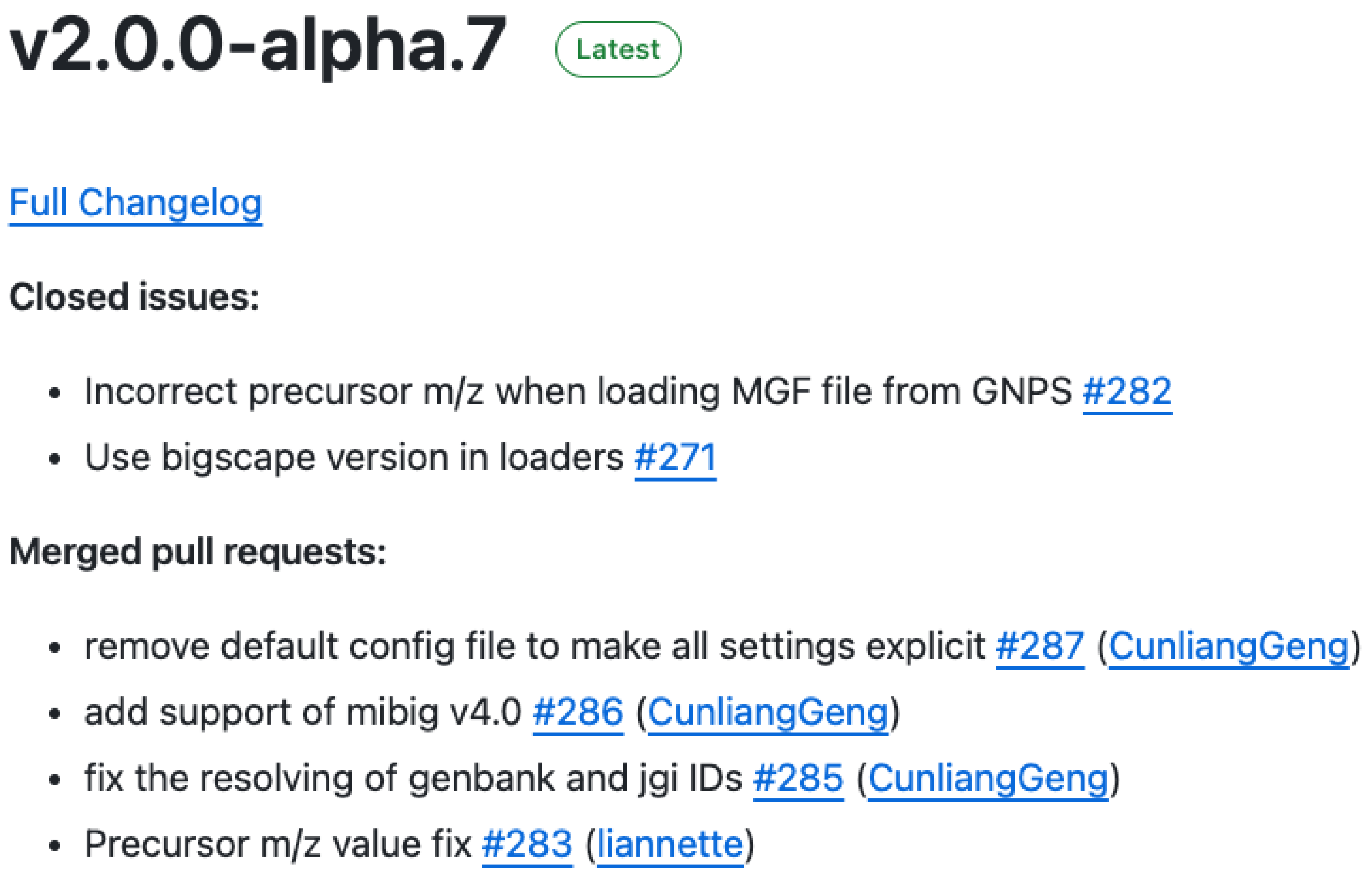
netherlands eScience center

Announcement & upcoming events

netherlands eScience center

NPLinker Development

New features & changes



v2.0.0-alpha.7

Full Changelog

Closed issues:

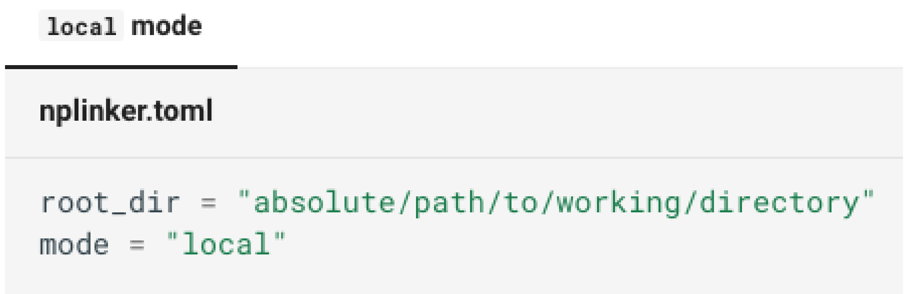
- Incorrect precursor m/z when loading MGF file from GNPS #282
- Use bigscape version in loaders #271

Merged pull requests:

- remove default config file to make all settings explicit #287 (CunliangGeng)
- add support of mibig v4.0 #286 (CunliangGeng)
- fix the resolving of genbank and jgi IDs #285 (CunliangGeng)
- Precursor m/z value fix #283 (liannette)

Now, all settings are configured by the user

Before 2.0.0-a7

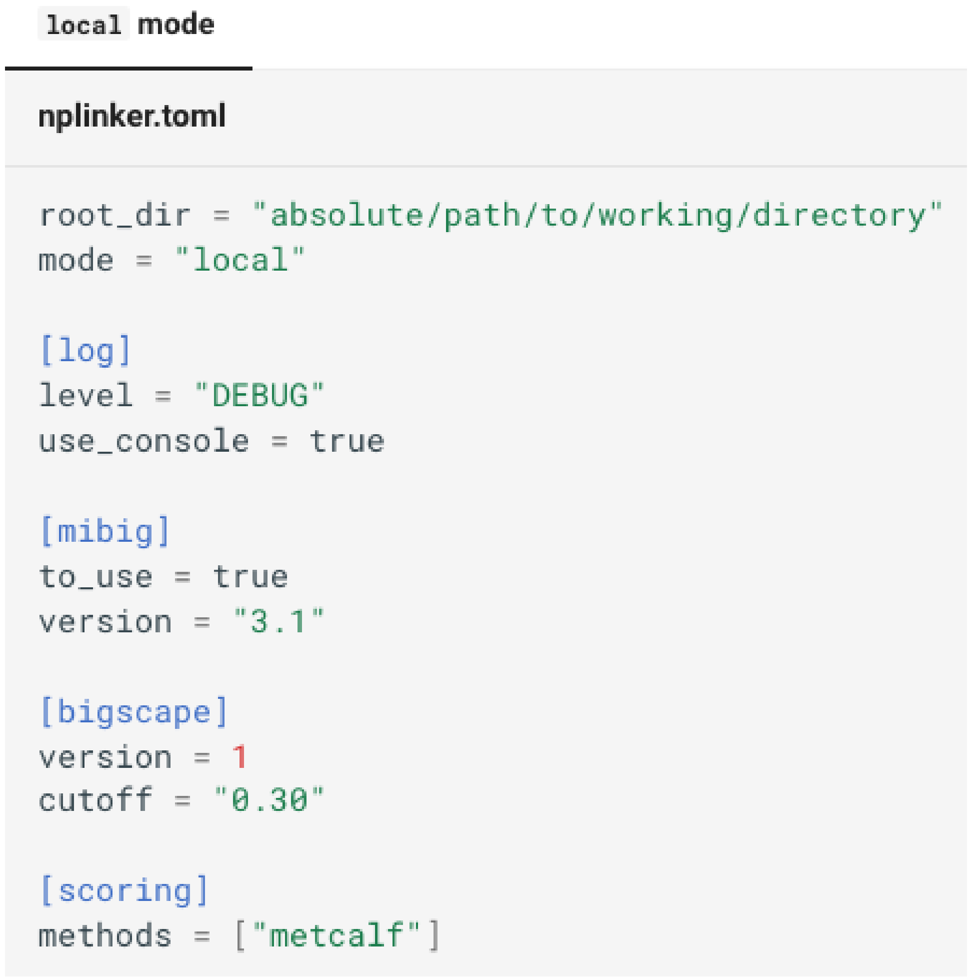


`local` mode

nplinker.toml

```toml
root_dir = "absolute/path/to/working/directory"
mode = "local"
```

Now



`local` mode

nplinker.toml

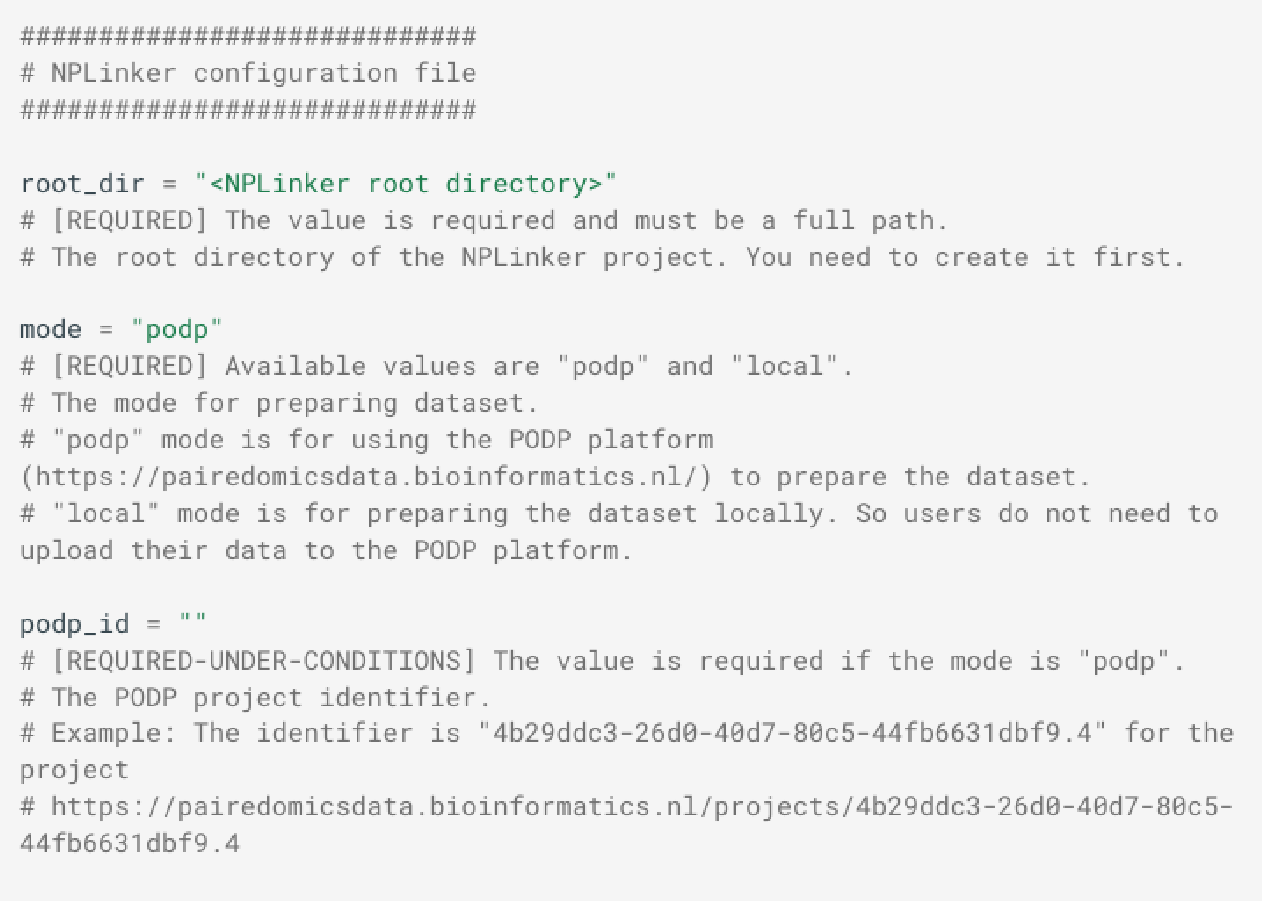
```toml
root_dir = "absolute/path/to/working/directory"
mode = "local"

[log]
level = "DEBUG"
use_console = true

[mibig]
to_use = true
version = "3.1"

[bigscape]
version = 1
cutoff = "0.30"

[scoring]
methods = ["metcalf"]
```

[PR 287]

Now, all settings are configured by the user

Check all required settings in the template file

https://nplinker.github.io/nplinker/latest/concepts/config_file/

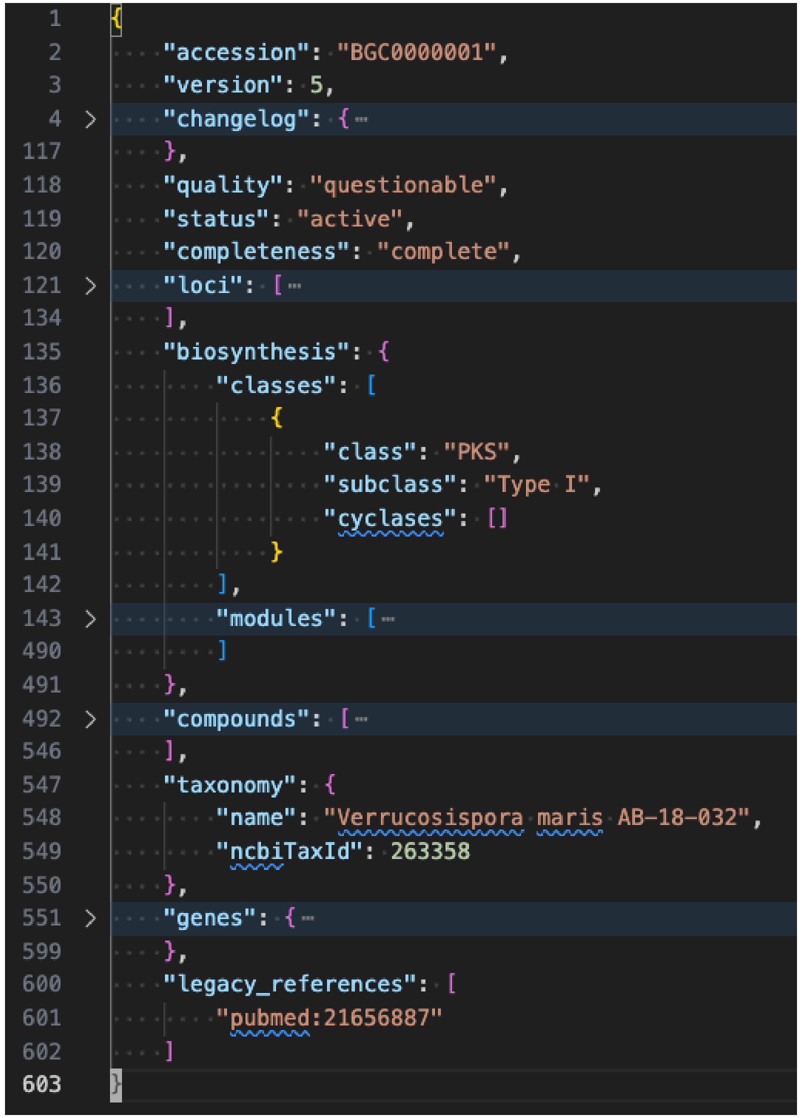


```
#############################
# NPLinker configuration file
#############################

root_dir = "<NPLinker root directory>"
# [REQUIRED] The value is required and must be a full path.
# The root directory of the NPLinker project. You need to create it first.

mode = "podp"
# [REQUIRED] Available values are "podp" and "local".
# The mode for preparing dataset.
# "podp" mode is for using the PODP platform
(https://pairedomicsdata.bioinformatics.nl/) to prepare the dataset.
# "local" mode is for preparing the dataset locally. So users do not need to
upload their data to the PODP platform.

podp_id = ""
# [REQUIRED-UNDER-CONDITIONS] The value is required if the mode is "podp".
# The PODP project identifier.
# Example: The identifier is "4b29ddc3-26d0-40d7-80c5-44fb6631dbf9.4" for the
project
# https://pairedomicsdata.bioinformatics.nl/projects/4b29ddc3-26d0-40d7-80c5-
44fb6631dbf9.4
```

[PR 287]

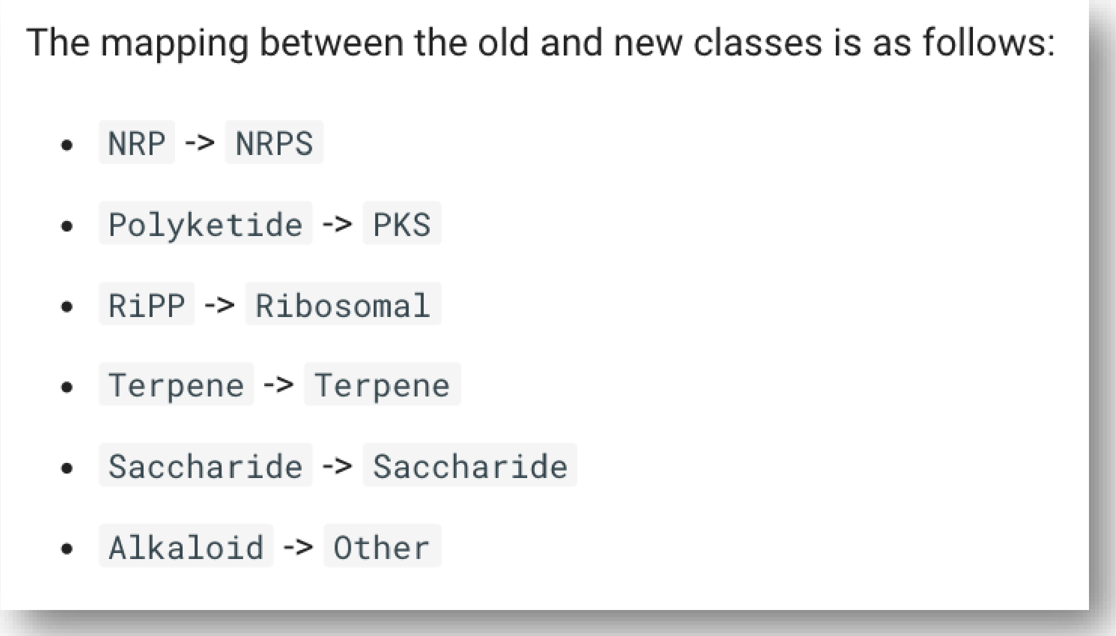
Support for MIBiG v4.0

MIBiG v4.0 uses a new schema for its metadata



```
{
    "accession": "BGC0000001",
    "version": 5,
    "changelog": {⋯
    },
    "quality": "questionable",
    "status": "active",
    "completeness": "complete",
    "loci": [⋯
    ],
    "biosynthesis": {
        "classes": [
            {
                "class": "PKS",
                "subclass": "Type I",
                "cyclases": []
            }
        ],
        "modules": [⋯
        ]
    },
    "compounds": [⋯
    ],
    "taxonomy": {
        "name": "Verrucosispora maris AB-18-032",
        "ncbiTaxId": 263358
    },
    "genes": {⋯
    },
    "legacy_references": [
        "pubmed:21656887"
    ]
}
```

Changes of biosynthetic classes



The mapping between the old and new classes is as follows:

- `NRP` -> `NRPS`
- `Polyketide` -> `PKS`
- `RiPP` -> `Ribosomal`
- `Terpene` -> `Terpene`
- `Saccharide` -> `Saccharide`
- `Alkaloid` -> `Other`

[PR 286]

Support for NCBI Datasets v2 REST API

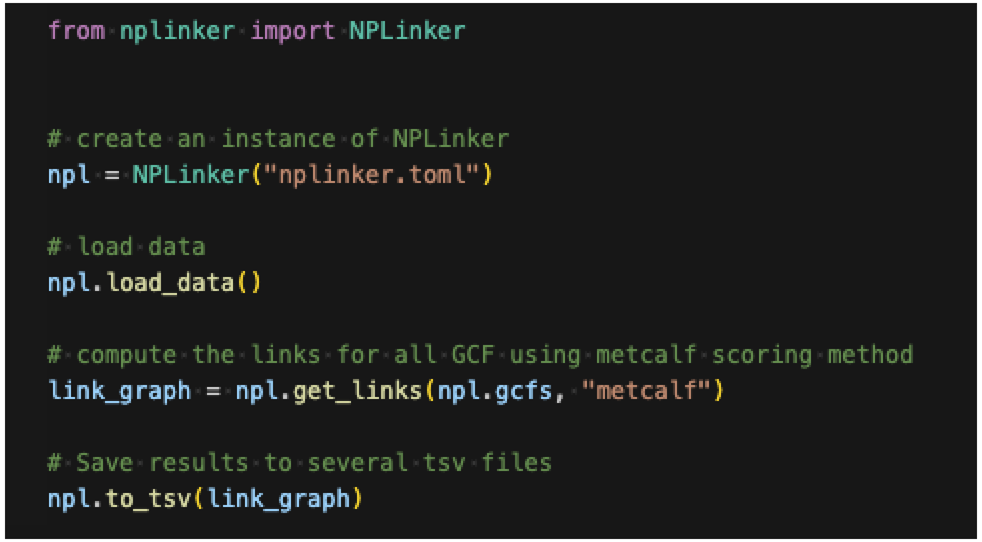
NCBI is required to resolve the RefSeq assembly accession in NPLinker.

**Previously NPLinker parses NCBI webpages to get the accession,
but now it uses the new NCBI REST API to do that.**

https://www.ncbi.nlm.nih.gov/datasets/docs/v2/api/rest-api/

[PR 285]

New: Save results in tabular output files



```python
from nplinker import NPLinker


# create an instance of NPLinker
npl = NPLinker("nplinker.toml")

# load data
npl.load_data()

# compute the links for all GCF using metcalf scoring method
link_graph = npl.get_links(npl.gcfs, "metcalf")

# Save results to several tsv files
npl.to_tsv(link_graph)
```

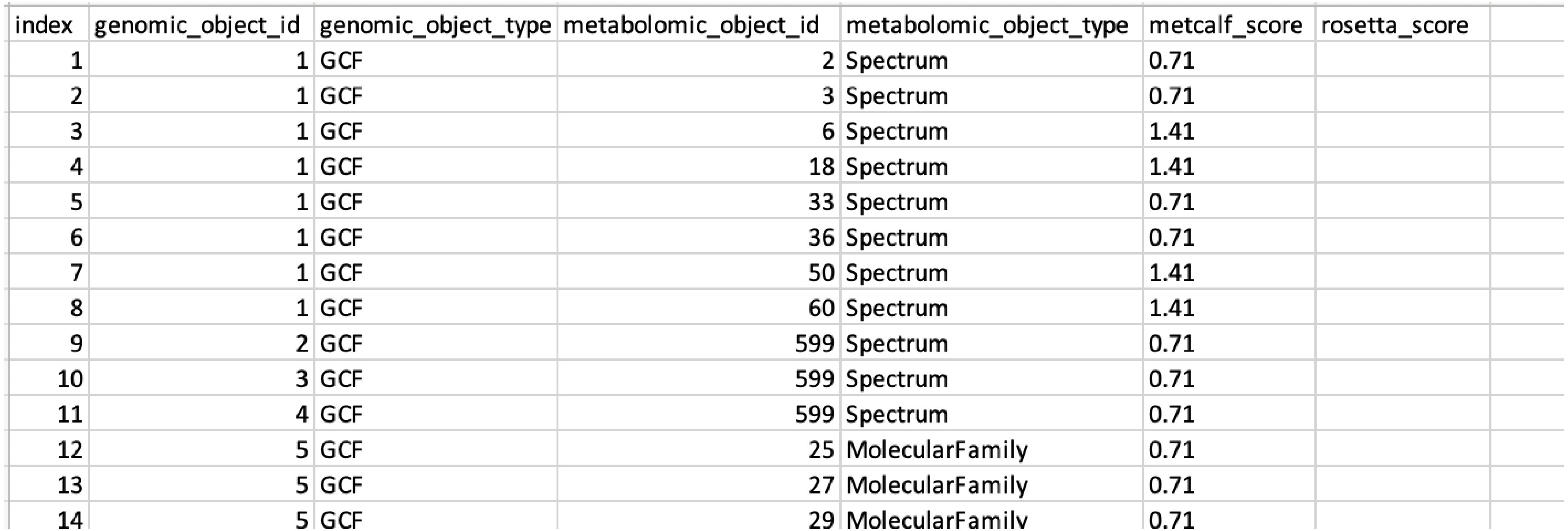
Export data to TSV files

This method exports all necessary data for further analysis to several TSV files:

• **BGCs**	**genomics_data.tsv**
• **Spectra**	**metabolomics_data.tsv**
• **Links**	**links.tsv**

New: Save results in tabular output files

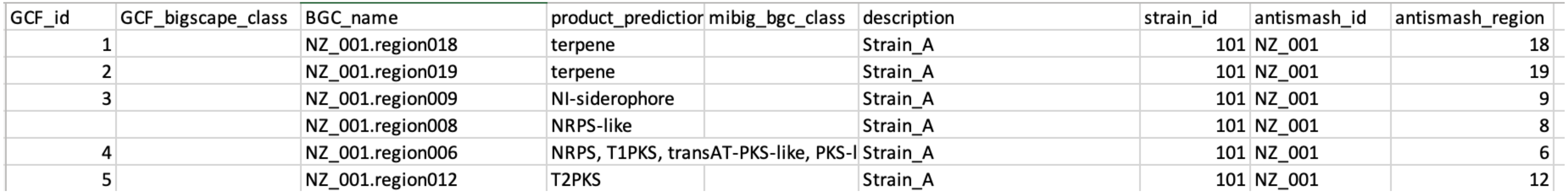
The calculated links between GCF and Spectra/Molecular Families are exported to the links.tsv file:



index	genomic_object_id	genomic_object_type	metabolomic_object_id	metabolomic_object_type	metcalf_score	rosetta_score
1	1	GCF	2	Spectrum	0.71	
2	1	GCF	3	Spectrum	0.71	
3	1	GCF	6	Spectrum	1.41	
4	1	GCF	18	Spectrum	1.41	
5	1	GCF	33	Spectrum	0.71	
6	1	GCF	36	Spectrum	0.71	
7	1	GCF	50	Spectrum	1.41	
8	1	GCF	60	Spectrum	1.41	
9	2	GCF	599	Spectrum	0.71	
10	3	GCF	599	Spectrum	0.71	
11	4	GCF	599	Spectrum	0.71	
12	5	GCF	25	MolecularFamily	0.71	
13	5	GCF	27	MolecularFamily	0.71	
14	5	GCF	29	MolecularFamily	0.71	

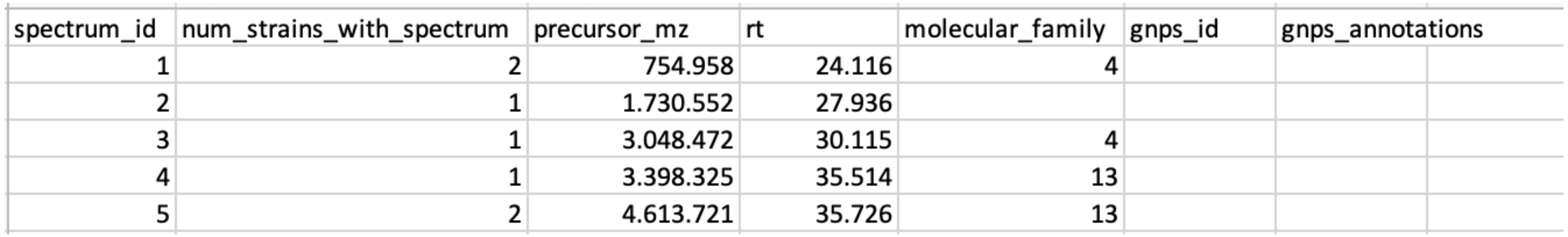
New: Save results in tabular output files

Information about the BGCs belonging to the GCFs are found in the genomics_data.tsv:



GCF_id	GCF_bigscape_class	BGC_name	product_predictior	mibig_bgc_class	description	strain_id	antismash_id	antismash_region
1		NZ_001.region018	terpene		Strain_A	101	NZ_001	18
2		NZ_001.region019	terpene		Strain_A	101	NZ_001	19
3		NZ_001.region009	NI-siderophore		Strain_A	101	NZ_001	9
		NZ_001.region008	NRPS-like		Strain_A	101	NZ_001	8
4		NZ_001.region006	NRPS, T1PKS, transAT-PKS-like, PKS-l		Strain_A	101	NZ_001	6
5		NZ_001.region012	T2PKS		Strain_A	101	NZ_001	12

Same as for the Spectra/MolecularFamilies in the metabolomics_data.tsv file:



spectrum_id	num_strains_with_spectrum	precursor_mz	rt	molecular_family	gnps_id	gnps_annotations
1	2	754.958	24.116	4		
2	1	1.730.552	27.936			
3	1	3.048.472	30.115	4		
4	1	3.398.325	35.514	13		
5	2	4.613.721	35.726	13		

netherlands eScience center

NPLinker Webapp Development