

NPLinker updates

Cunliang Geng & Giulia Crocioni

2024-02-09

netherlands
eScience center

Starting point

A public **software repo** with a published paper

nplinker

Public

Edit Pins

Unwatch 6

release-1.2.0

5 Branches

12 Tags

Go to file

Add file

Code

This branch is 322 commits behind dev

Contribute

andrewramsay

Add missing date for 1.2.0 release

d4aae71 · 2 years ago

973 Commits

.vscode

changed handling of known cluster blast results and adde...

6 years ago

docs

add sphinx master_doc

4 years ago

notebooks

Expand explanations

2 years ago

prototype

Merge pull request #65 from louwenjir/npclassscore_doc...

2 years ago

webapp/npapp

remove old comments and imports

2 years ago

.dockerignore

update dockerignore

4 years ago

COMPUTATIONAL BIOLOGY

advanced search

Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions

Grímur Hjörleifsson Eldjárn, Andrew Ramsay, Justin J. J. van der Hooft, Katherine R. Duncan, Sylvia Soldatou, Juho Rousu, Rónán Daly, Joe Wandy, Simon Rogers

Version 2 Published: May 4, 2021 • <https://doi.org/10.1371/journal.pcbi.1008920>

43 Save	9 Citation
2,469 View	24 Share

<https://doi.org/10.1371/journal.pcbi.1008920>

Challenges we have faced from the beginning

Creating a Scalable NPLinker
Framework with Interactive
Visualization Module

Code reusability challenge

Need a more user-friendly web app



Redesign & Refactor (Rewrite)



The principles of refactoring

- Easy to install & run 🙌 packing code to a package/image, having tutorials and documentations, ...
- Easy to read 🙌 clear structure of codebase, proper formatting of code, ...
- Easy to understand 🙌 meaningful names, static typing, good comments, ...
- Easy to change 🙌 well-designed architecture (abstractions), ...
- Modular code 🙌 use function and class to organise code, do one thing per function/class,...
- Correct code 🙌 unit test, integration test, user test...
- Don't repeat yourself 🙌 eliminating duplicated code
- Don't reinvent the wheel 🙌 taking advantage of existing libraries and packages
- ...



What we have done

- Split the software repo into two repos: [nplinker](#) and [webapp](#)
- Created python package for nplinker (v1.3.2) and simplified the installation
- Redesigned the architecture of nplinker into two main parts: data preparation and scoring
- Ongoing refactoring of the data preparation part
- Redesigned the UI of webapp and explored tech solutions
- Got an internal grant (0.5fte) to improve the sustainability of NPLinker software

Code changes in 2023

Showing changes from 194 commits

± 197 changed files +11930 -9986 ⚙️ ▼

.github/workflows

70 added

.github/workflows

86 modified

.gitignore

25 deleted

16 renamed



Research Software Engineer
Dr. Cunliang Geng



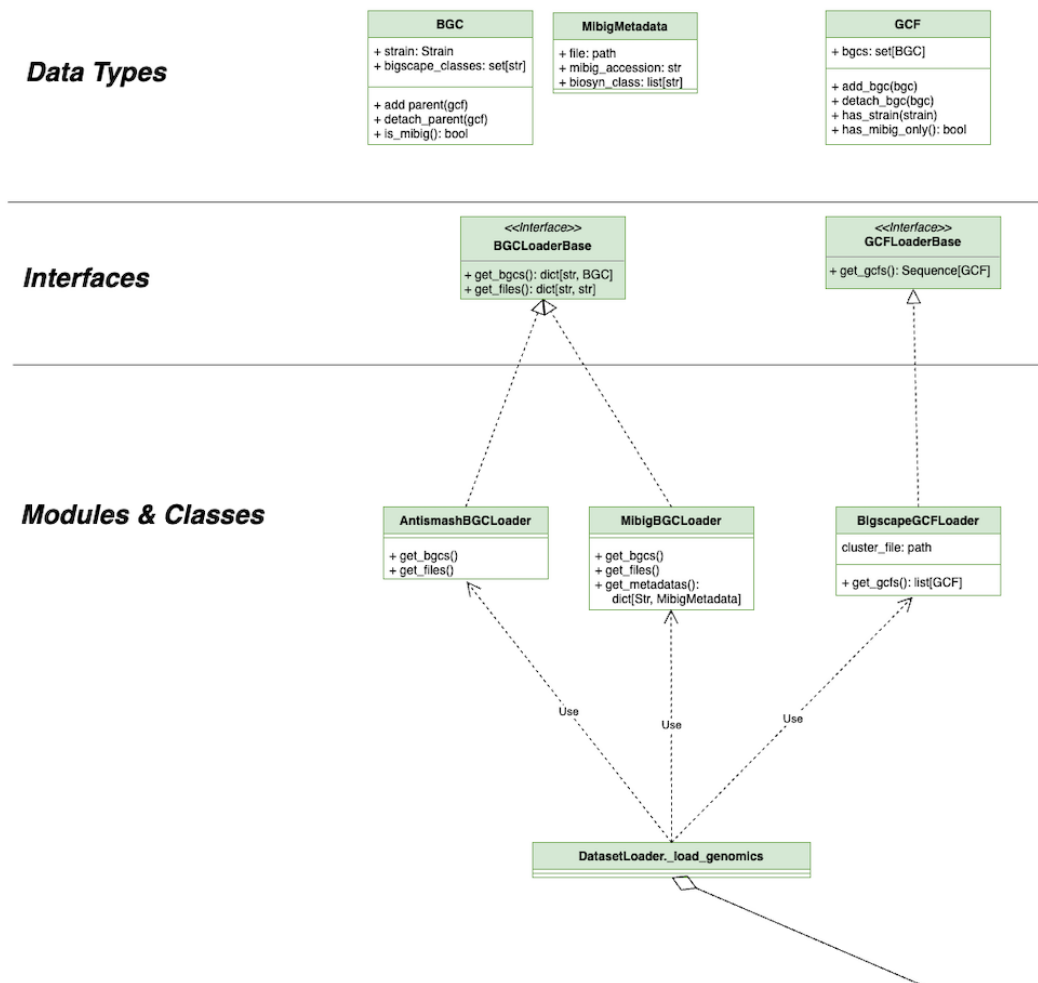
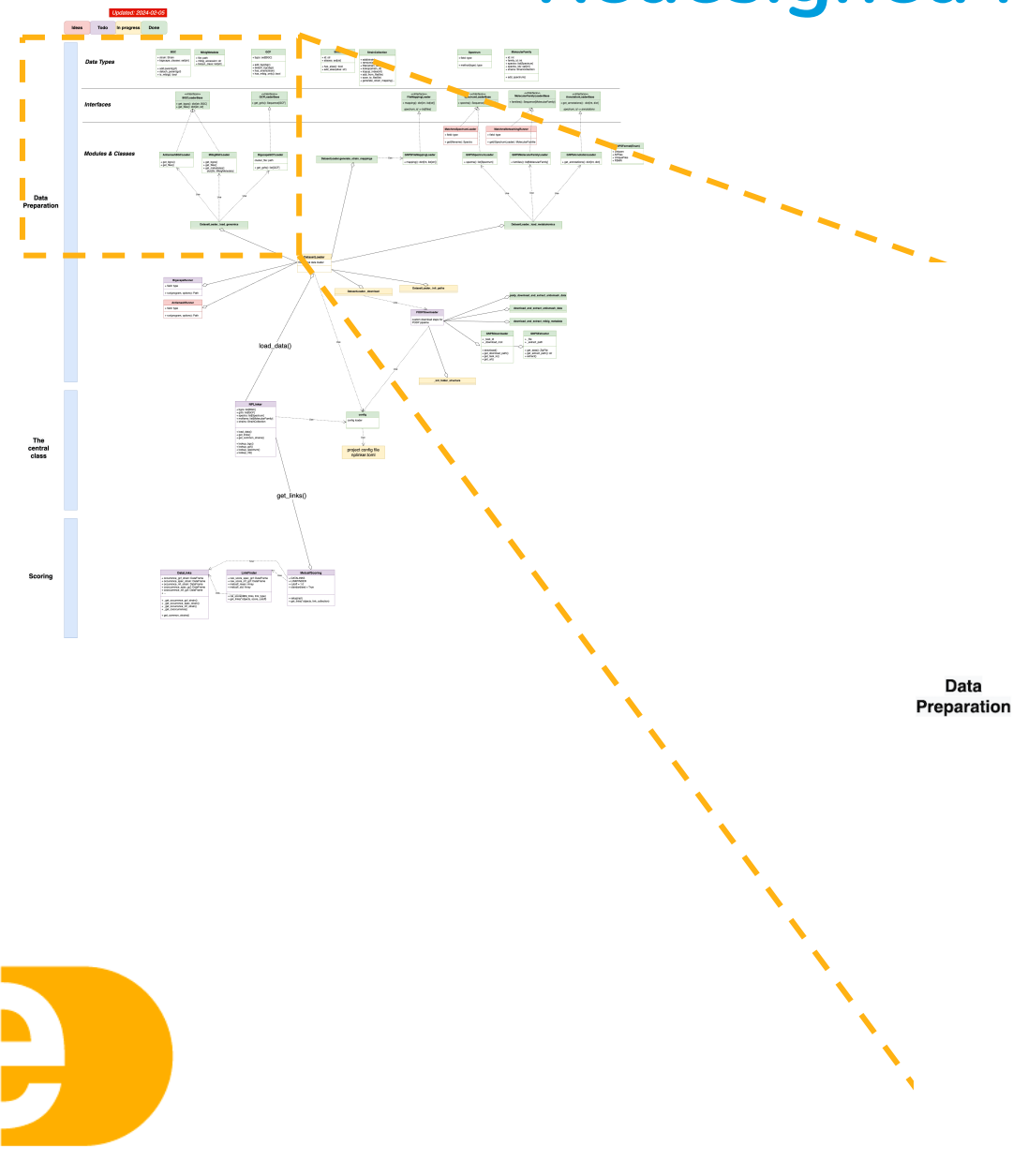
Research Software Engineer
Giulia Crocioni, MSc



PhD candidate
Helge Hecht, MSc



* Not all functions/classes displayed



Data loading pipeline

Updated: 2023-02-05
Author: Cunliang Geng

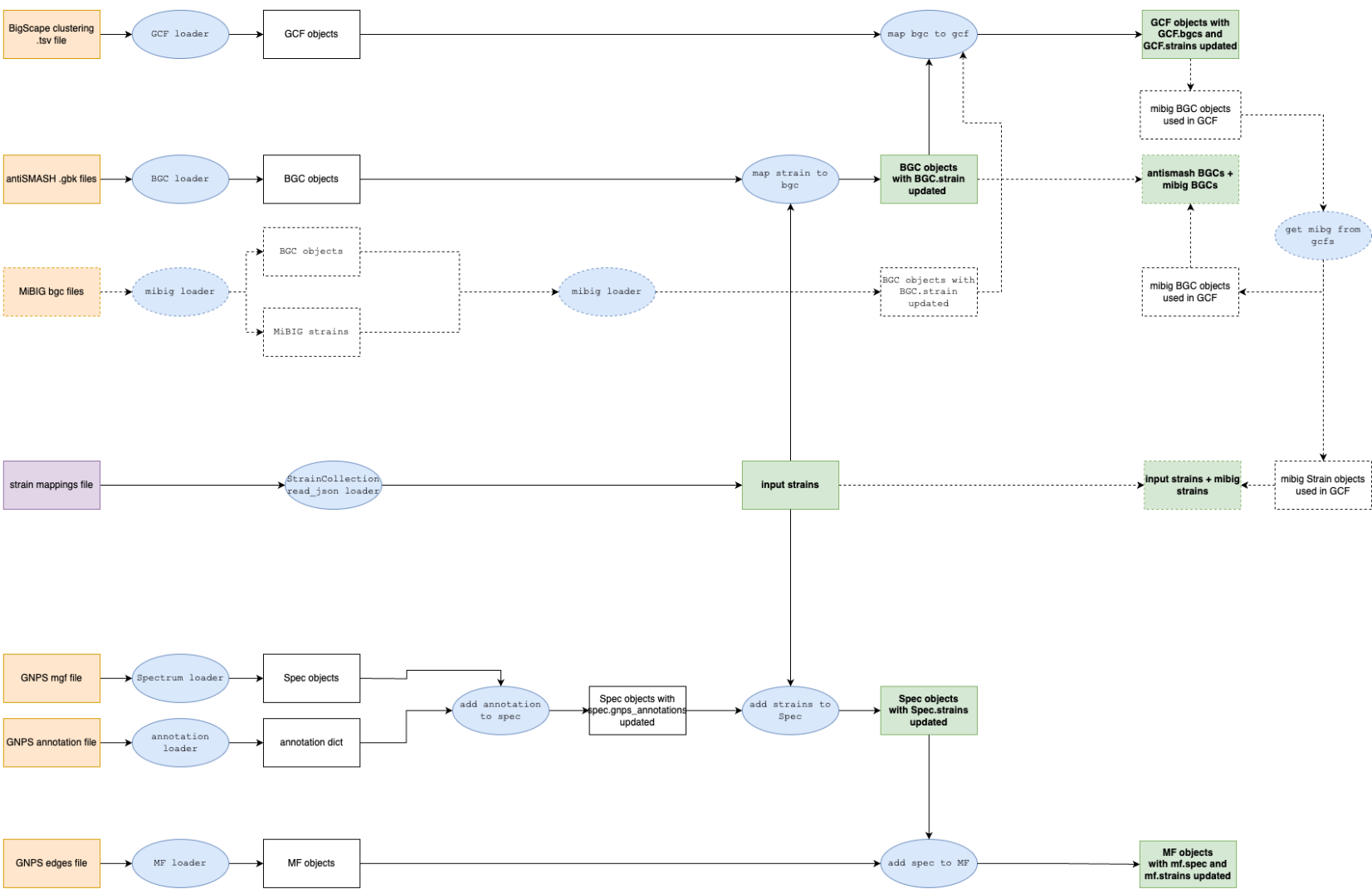
input or output

function

Genomics
load pipeline

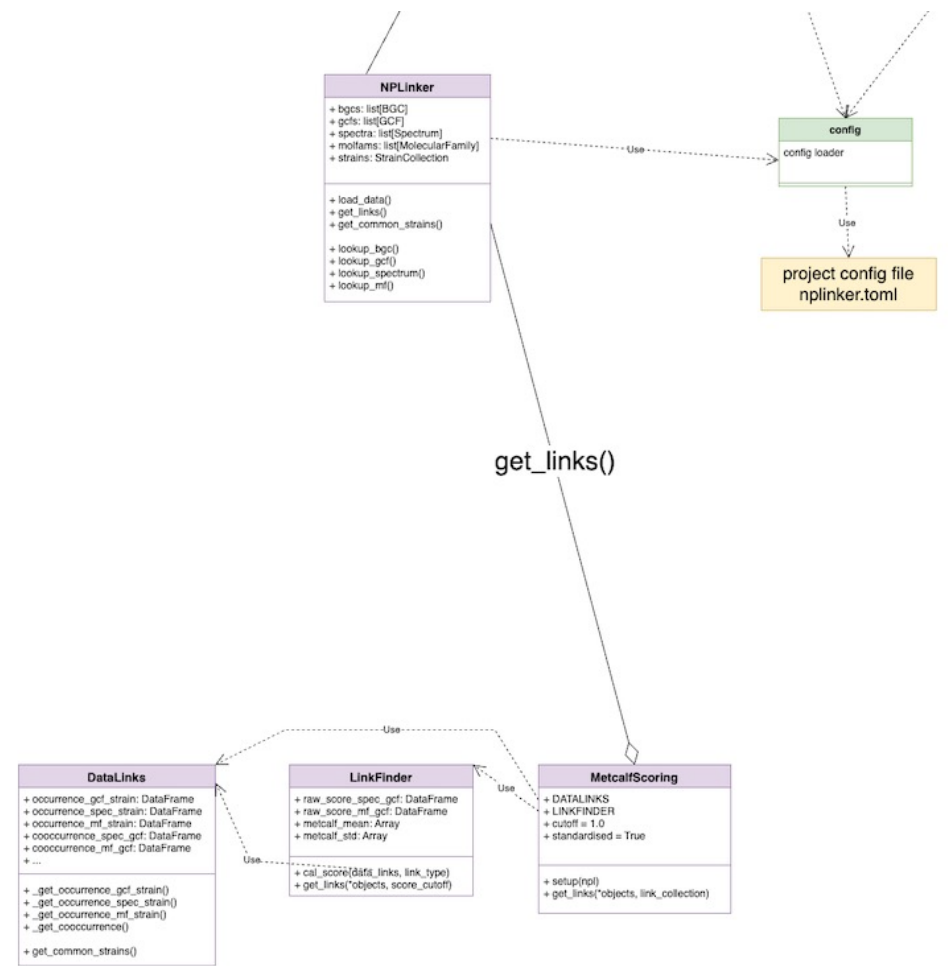
Strains
load pipeline

Metabolomics
load pipeline



Redesigned NPLinker architecture

* Not all functions/classes displayed



Some highlights of the refactoring

- Restructured the codebase and uniformed the formatting of code with ruff
- Added docstrings
- Added python static typing
- Added unit tests and parallel testing
- Added/updated code for data objects, downloaders, extractors and loaders
- Updated the logics of loading GEN and MET data, making them consistent with each other
- Updated the logics of loading strain mappings, making it decoupled from loading of GEN/MET data
- Changed the output file format (csv/tsv) to JSON
- Added schema for input/output JSON files and added schema validators
- Updated the loading of config with config manager Dynaconf
- ...



Redesigned Webapp UI

NPLinker 1 Import data Import Doc About

2 Choose the tab according to the data you want to start with, e.g. Genomics data

3 Filter data using filters (optional)

Genomics filter: GCF ID, Type, Foldable tab

GCF ID: Enter one or more GCF IDs

Type: Enter one or more compound types

4b Inspect the data (optional)

h1000

4b Select data

Data: 2 GCFs selected

GCF ID	# BGC
1	3
2	1
3	18
4	6

Rows per page: 20 <>

5 Set scoring methods and parameters (optional)

Metacalif: RAW, STANDARDIZED

Outoff: 1

Spectral score outoff: 0

BGC score outoff: 0

Logical operation between scoring methods: AND, OR

6 Click to show the links found

Show spectra

7 Inspect the results

Candidate Links

GCF ID	# links	Top 1 spectrum	Product	Score
1	300	100	xxx	0.8
4	20	20	yyy	0.8
2	400	45	zzz	0.6

Rows per page: 3 <>

8 GCF Viewer - When click on over

GCF 1

GCF ID: 1

Class: PFB-BSP_Hybride

Strains: Salinispora arenicola

CSQ748

BigFam ID: none (if exists)

BGC name	Product	Hybrid?
BGC_1		
BGC_2		
BGC_3		

9 Spectrum Viewer - When click on over

Spectrum 100

Spectrum ID: 100

Predecessor ms: 718.28

Parent ms: 717.27

Strains: Salinispora arenicola

GNPS hit: link (if exists)

10 Scores

Merged score: 0.6

Metacalif score: 0.45

Rosetta score: 0.52

Also HPCClassScore will be here after having added it to the code-base

There will be a ranking algorithm for merging different scores values.

NPLinker 1 Import data Import Doc About

2 Choose the tab according to the data you want to start with, e.g. Metabolomics data

3 Filter data using filters (optional)

Metabolomics filter: MF ID, Spectrum ID

MF ID: Enter one or more Molecular Family IDs

Spectrum ID: Enter one or more Spectrum IDs

4 Select data

Data: 2 MFs selected

MF ID	Spectrum ID
1	45
2	2
3	2
4	47

Rows per page: 20 <>

5 Set scoring methods and parameters (optional)

Metacalif: RAW, STANDARDIZED

Outoff: 1

Spectral score outoff: 0

BGC score outoff: 0

Logical operation between scoring methods: AND, OR

6 Click to show the links found

Show GCFs

7 Inspect the results

Candidate Links

MF ID	Spectrum ID	# links	Top 1 GCF	Product	Score
1	45	14	4	xxx	0.8
12	4	20	58	yyy	0.8
4	47	48	2	zzz	0.6

Rows per page: 3 <>

8 Spectrum Viewer

Spectrum 45

Spectrum ID: 45

Predecessor ms: 718.28

Parent ms: 717.27

Strains: Salinispora arenicola

GNPS hit: link (if exists)

9 GCF Viewer

GCF 4

GCF ID: 4

Class: PFB-BSP_Hybride

Strains: Salinispora arenicola

CSQ748

BigFam ID: none (if exists)

BGC name	Product	Hybrid?
BGC_1		
BGC_2		
BGC_3		

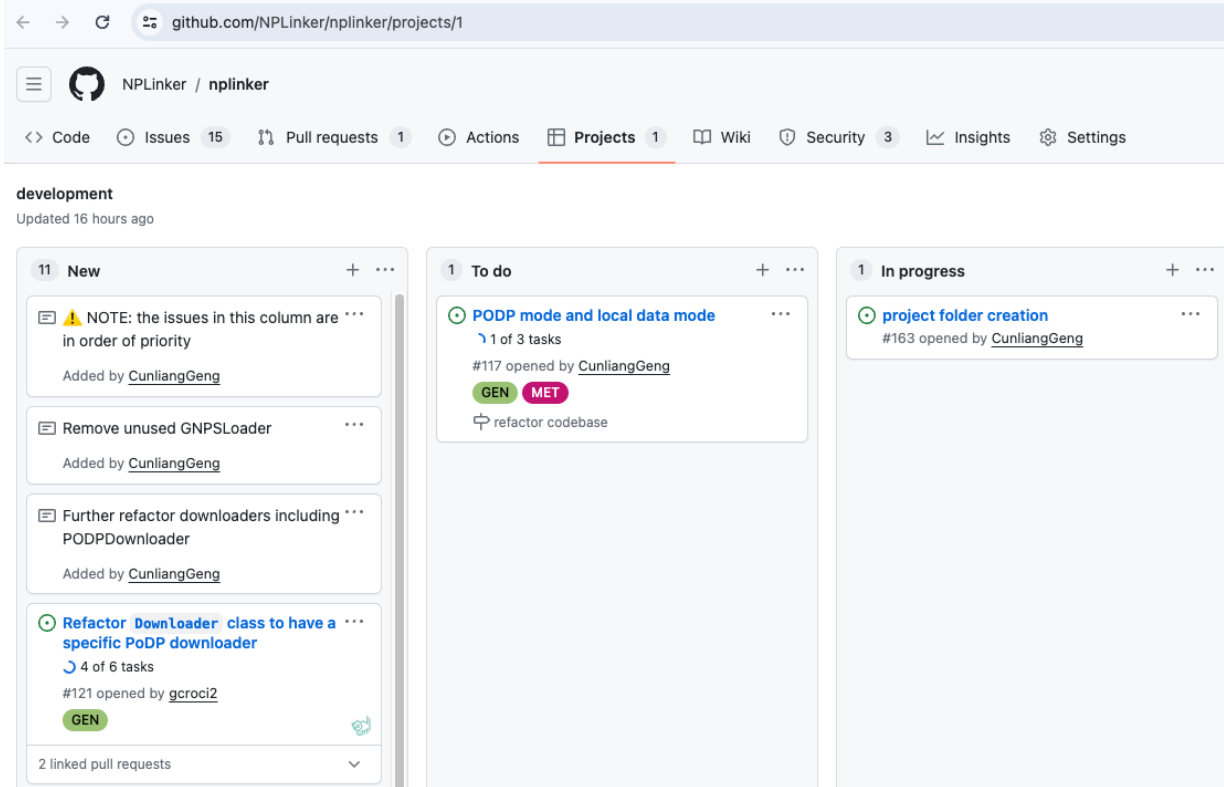
10 Scores

Merged score: 0.8

Metacalif score: 0.84

Rosetta score: 0.79

Future work



Kanban

<https://github.com/NPLinker/nplinker/projects/1>

Future work

- Enable the pipeline of loading user's local data (Local mode)
- Update template of config file and the creation of project directory structure
- Write tutorials for user testing on loading data
- Develop webapp
- Refactor scoring part
- Further user testing



Let's stay in touch

 www.eScienceCenter.nl

 c.geng@esciencecenter.nl

 [@CunliangGeng](https://github.com/CunliangGeng)