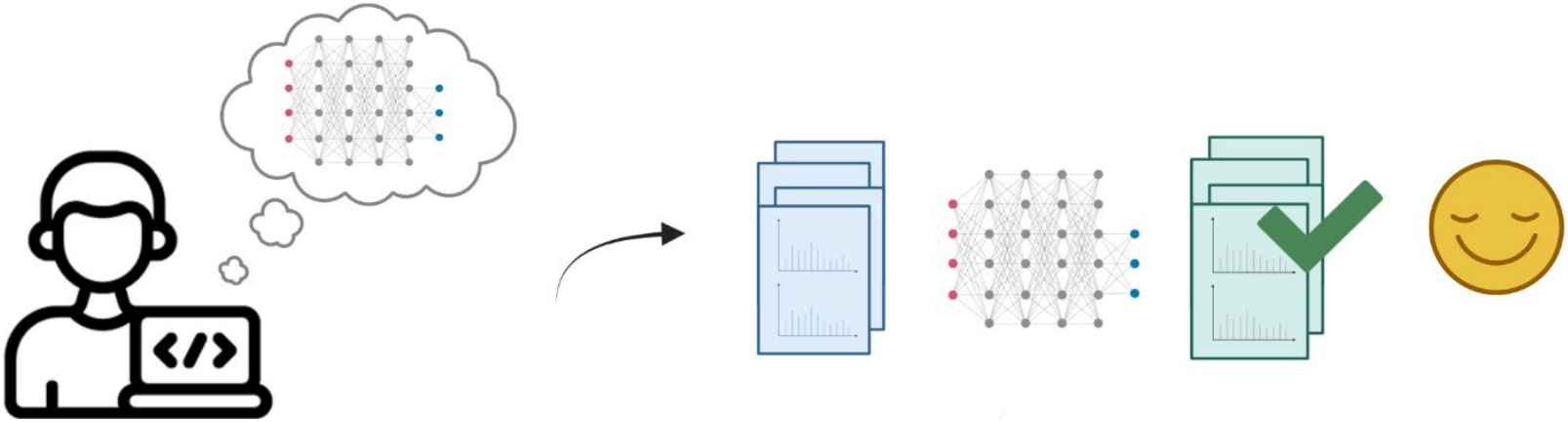
Mass Dynamics

Automated benchmarking enables continuous confidence in scientific software development

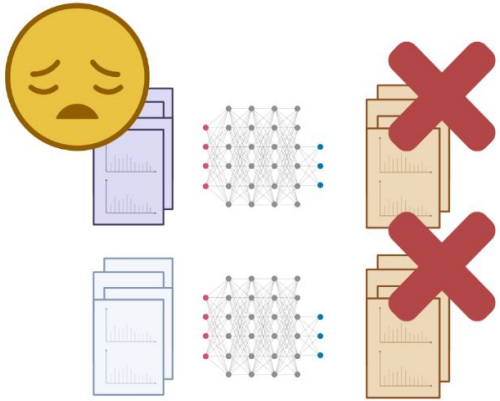
Anna Quaglieri, PhD
Data Scientist at Mass Dynamics

ASMS Minneapolis
7th June 2022



.....

6 months later



>8 years in Bioinformatics

Many hours spent looking for mysterious errors

- BSc, MSc in Statistics
- PhD in Statistical Genomics, WEHI Melbourne
- Since 1 year Data Scientist at Mass Dynamics, Melbourne

nature > nature human behaviour > perspectives > article

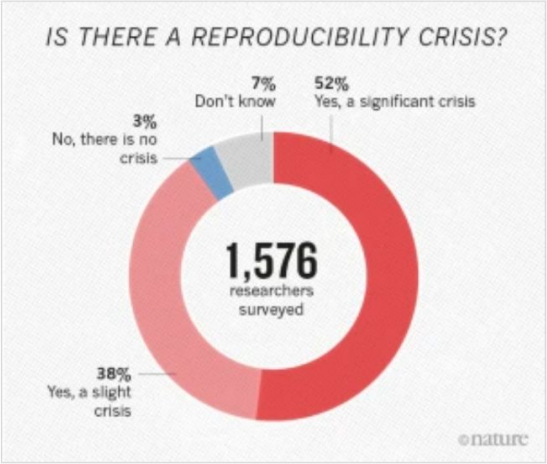
Open Access | Published: 10 January 2017

A manifesto for reproducible science

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

*"… one analysis estimates that **85% of biomedical research efforts are wasted**"*



Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

Nature **533**, 452–454 (2016) | Cite this article

What I would like you to get from my talk

What I would like you to get from my talk

- How at Mass Dynamics we get **quick** and **regular confidence** about the scientific software that we develop

What I would like you to get from my talk

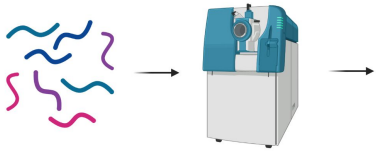
- How at Mass Dynamics we get **quick** and **regular confidence** about the scientific software that we develop

- **Learn ideas and methods** that you can apply to your own scientific development or processes

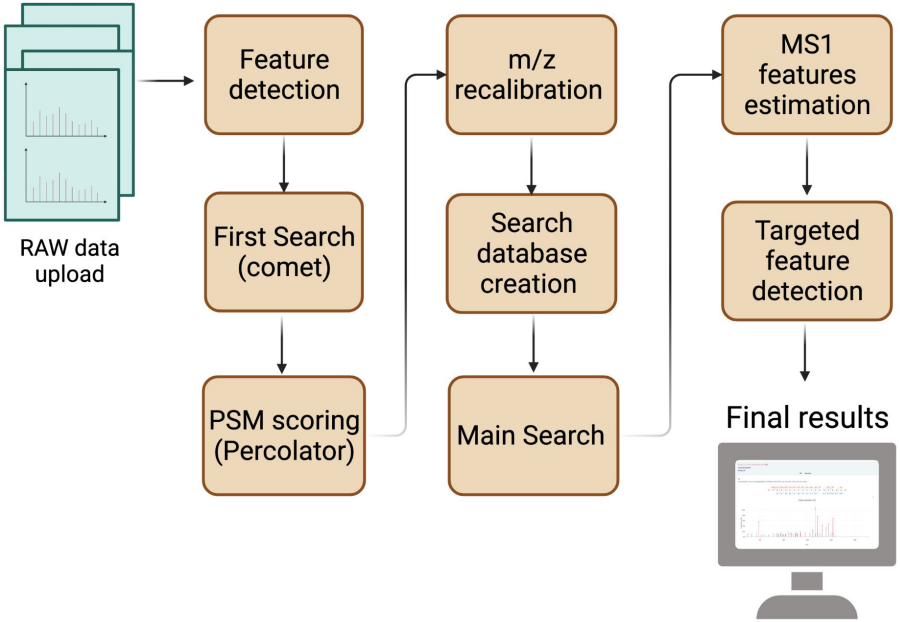
The problem in the proteomics context

Workflows in proteomics are often **Complex** and **computationally intensive**

Before Getting to Mass Dynamics



Example: Mass Dynamics Peptide Mapping workflow

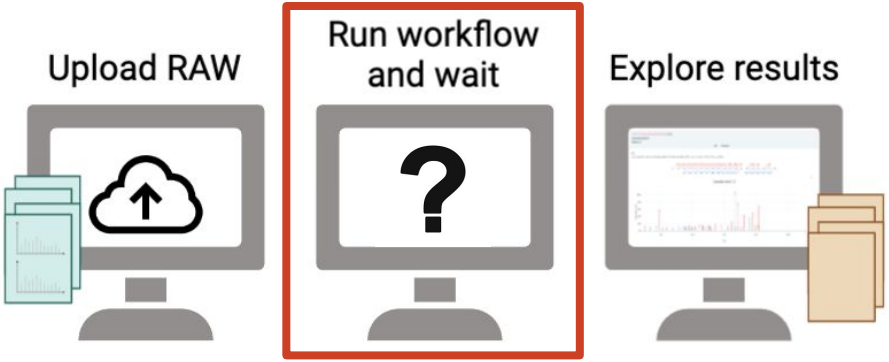


Automation overcomes complexity

Automation overcomes complexity

A complex automated workflow can feel like a black box for:

- Who uses it
- Who develops it



Benchmarking to maintain confidence

- Against ground truth before and after making changes to the code

Benchmarking to maintain confidence

- Against ground truth before and after making changes to the code
- Usually time consuming and performed manually
  - Benchmarking scripts across different folders and projects
  - To be run manually every time there is a new change

Imagine if…

We could all have a **quick development feedback**…

Imagine if…

We could all have a **quick development feedback**…

To maintain results **accuracy** after changes

Minimise risk of "silent" errors

Imagine if…

We could all have a **quick development feedback**…

To maintain results **accuracy** after changes

& to maintain **transparency** of internal steps

Build trust in the black box

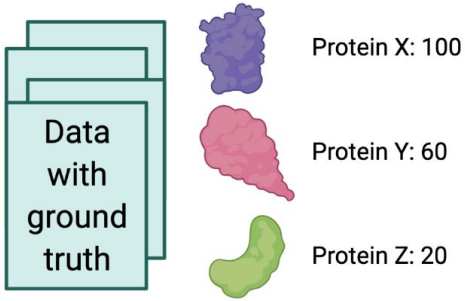
Our solution:
Automated Benchmarking System

Our solution: **Automated Benchmarking System**

Brings the CI/CD framework from software into scientific development

Produces automated reports to increase visibility of intermediate steps

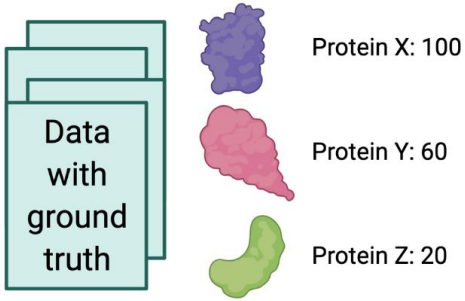
Benchmarking setup

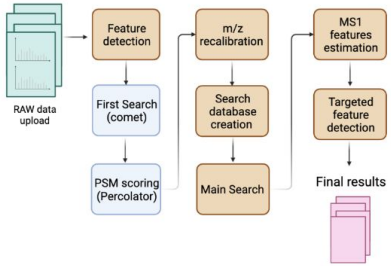


1 Data with known amounts of proteins

Public and private datasets

Benchmarking setup



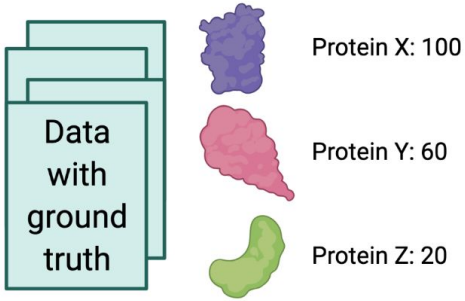


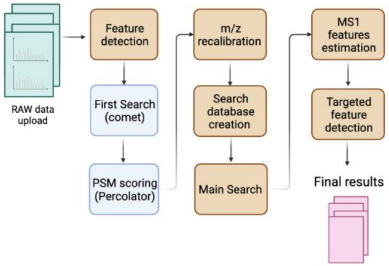
1 Data with known amounts of proteins

Public and private datasets

2 Workflow to run

Benchmarking setup





1 Data with known amounts of proteins

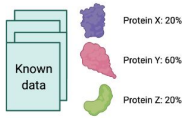
Public and private datasets

2 Workflow to run

3 Code to check results

Automated benchmarking system

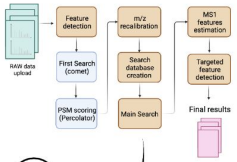
Data with ground truth



aws

1

Workflow



aws

2

Buildkite

3

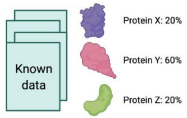
Centralised checks

GitHub

Created with BioRender.com

Automated benchmarking system

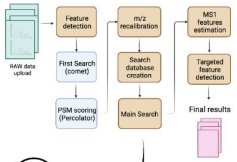
Data with ground truth



aws

1

Workflow



aws

2

Buildkite

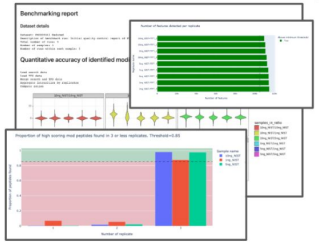
3

Centralised checks

GitHub

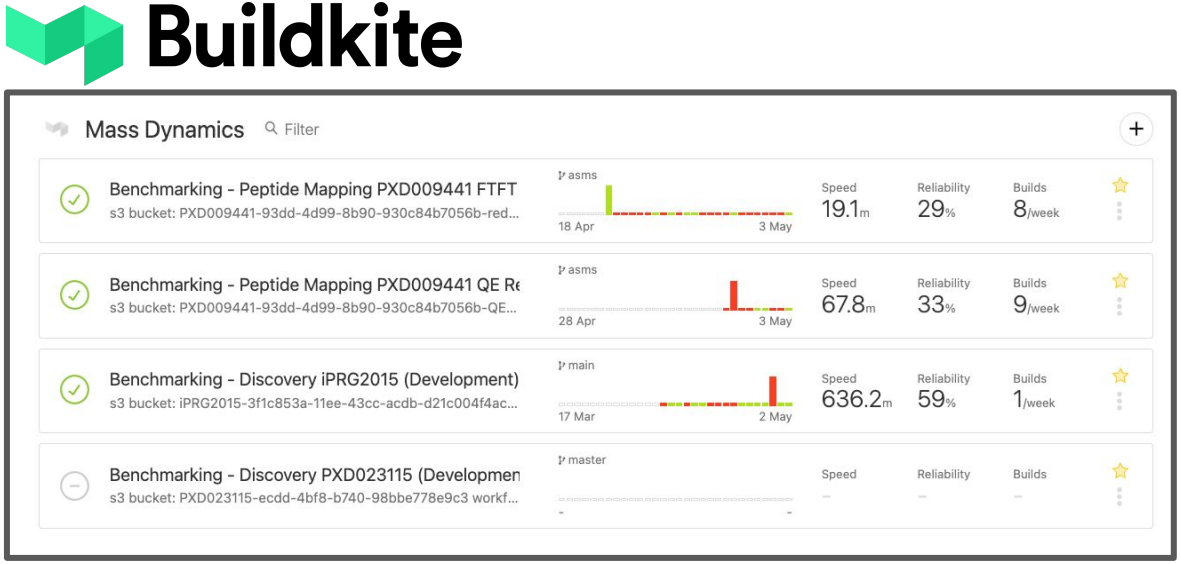
Automated tests and reports

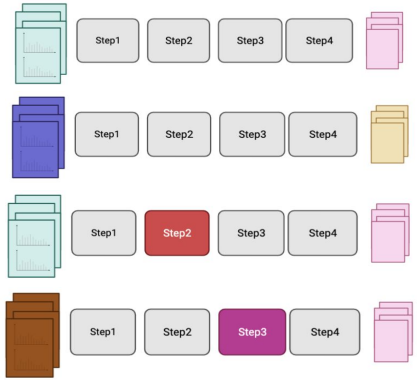
4



Created with BioRender.com

Each **dataset + workflow** has a separate pipeline

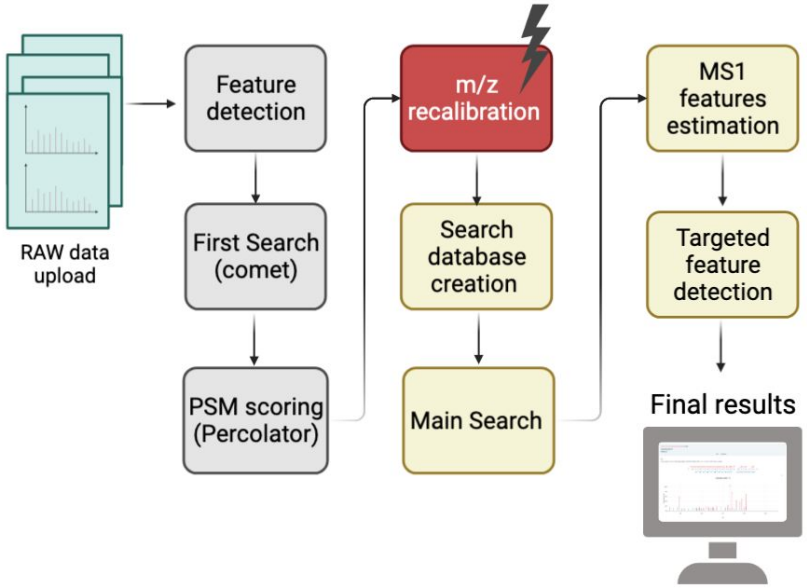




Created with BioRender.com

Automated benchmarking in action

After we make a change to a step: the underlying code changes



Created with BioRender.com

Create a new *build* for one pipeline with one click

Mass Dynamics / Benchmarking - Peptide Mapping PXD009441 / main
s3 bucket: PXD009441-93dd-4d99-8b90-930c84b7056b workflow: dcdfb159-d30c-4535-bf31...

All Builds | Edit Steps | Pipeline Settings | New Build

Test New Workflow
Build #98 | main | aa47550

Passed in 10s and blocked

buildkite-agent pi... | Be | Prepare python environ...

Remove deletable tags | Print | Rep...

Anna Quaglieri
Created today at 10:25 | Triggered

Rebuild | Cancel

buildkite-agent pipeline upl | Waited 6s | ip-172-31-18-181.ap-southeas...

New Build

Message

Test New Workflow

Description of the build. If left blank, the commit message will be used once the build starts.

Commit — Required

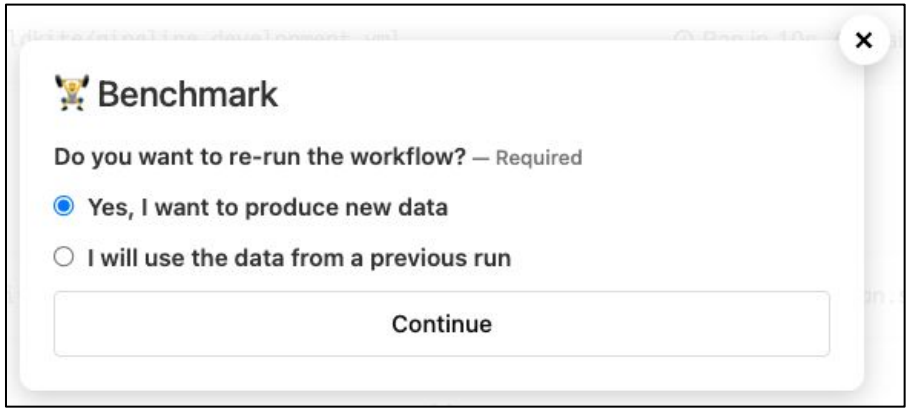
HEAD

Branch — Required

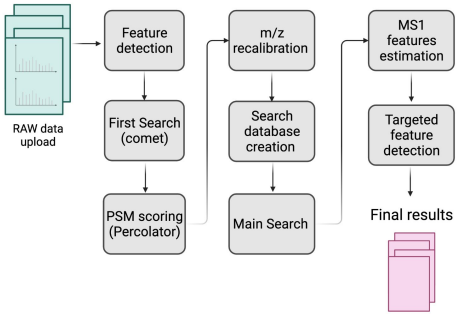
main

Options

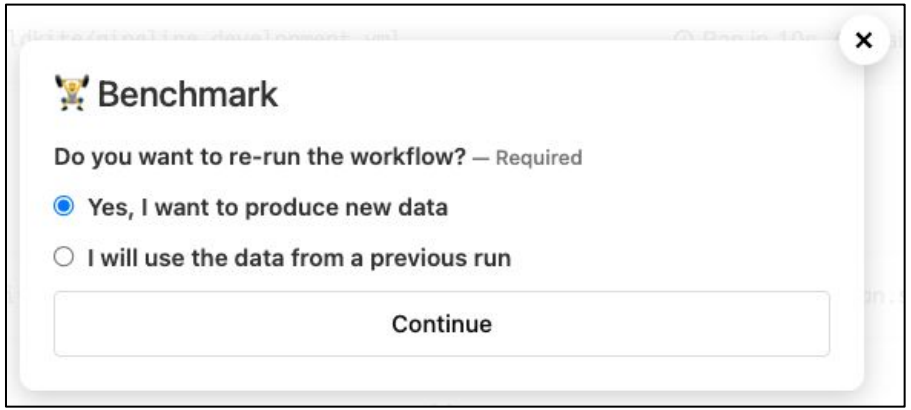
Create Build



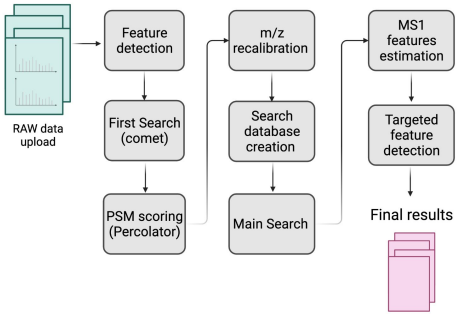
Run the Peptide Mapping Workflow end to end



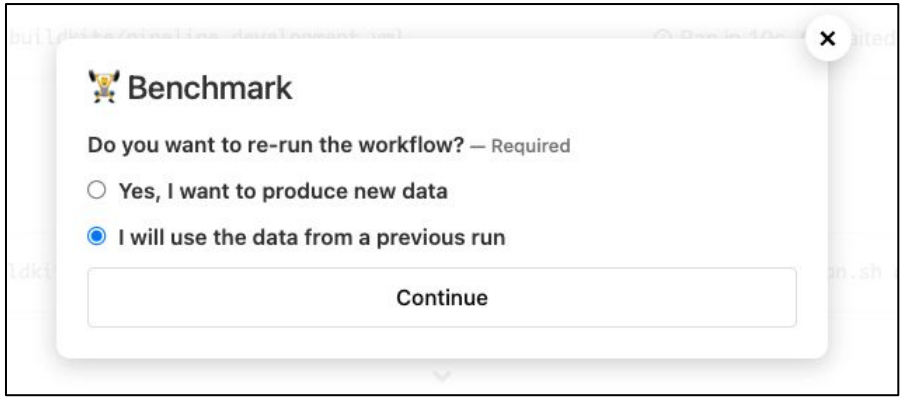
~ 2-7h



Run the Peptide Mapping Workflow end to end



~ 2-7h



Only run benchmarking code with the data already created

~ 10-15 min

If everything is good 

Mass Dynamics / Benchmarking - Peptide Mapping PXD009441 FTFT Reduce...
s3 bucket: PXD009441-93dd-4d99-8b90-930c84b7056b-reduced workflow: dcdfb159-d30c-4...

All Builds | Edit Steps | Pipeline Settings | New Build

Fix pipeline take 3
Build #27 | asms | a646cd6

Passed in 13m 3s

buildkite-agent pi... > Benchmark > Clean previous workflow > Start workflow > Wait for workflow > Prepare python environ... >
Remove deletable tags > Print AMIs > Test benchmarking code > Download + Create Rep... > Download + Test

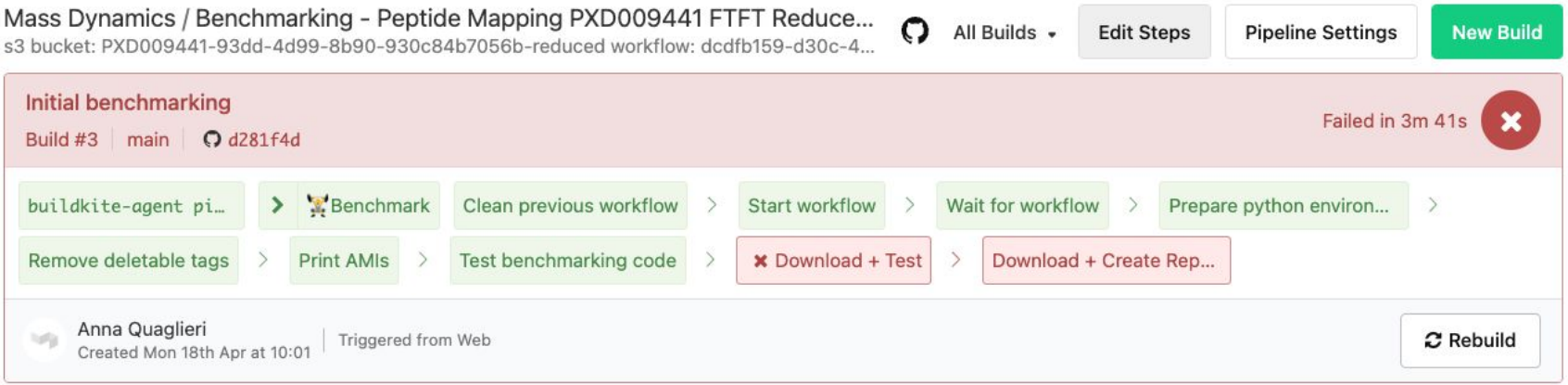
Anna Quaglieri
Created yesterday at 17:42 | Triggered from Web

Rebuild

buildkite-agent pipeline upload .buildkite/pipeline.development.yml

Ran in 6s | Waited 4s | ip-172-31-1-98.ap-southeast-...

If some tests fail

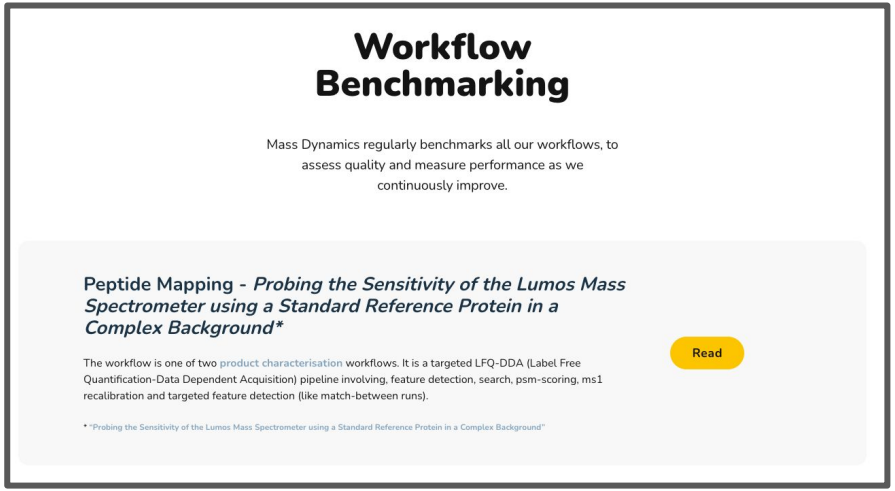


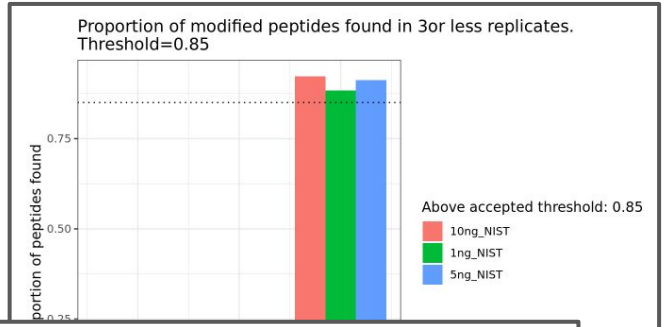
Public benchmarking report

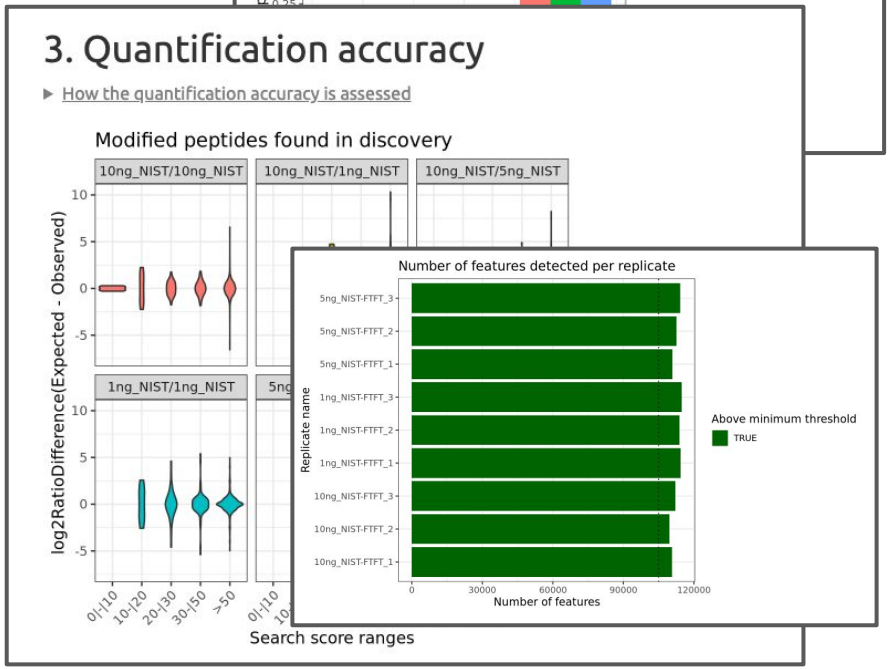
app.massdynamics.com/content/workflow/reports

Dataset with ground truth: PXD009441

Levy MG et al, 2018, JPR





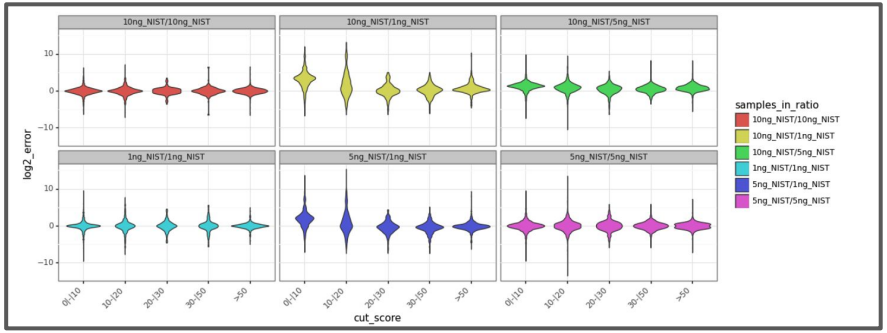


How we use the benchmarking system

Once we add a new dataset we use the results as our first benchmark

```yaml
name: PXD009441
description-benchmark: "Initial quality control report of PXD900441"
tests:
  data:
    number-of-runs: 24
    number-of-samples: 8
    replicates-per-sample: 3
  feature-detection-thermo:
    minimum-features-per-run: 100000
  protein-inference:
    expected-target-above-threshold: 0.001
    expected-decoy-below-threshold: 0.01
    q-value-threshold-psms: 0.01
    q-value-threshold-peptides: 0.01
```

Quantitative accuracy against ground truth



How we use the benchmarking setup

When the **workflow changes** we compare the results against the benchmark:

Did we improve?

Did we make it worse?

What changed?

For example: If we get a loss of identifications,

what has also changed in previous steps?

Today vs 1 year ago

- Gained quick and efficient way of getting confidence in our results
- We are more able to work concurrently than before
- Largely increased the spread of datasets that we check
- Gained a lot of knowledge about the intermediate steps and what to expect

Summary

Described the setup of Mass Dynamics **automated benchmarking framework**

To elevate the scientific development & benchmarking to the same level as the software development

Wins 🏆

- **Confidence in results Accuracy**:
  - We can make small or large changes safely
- **Transparency**:
  - Standardised quality control reports shine lights on hidden steps
  - Quality control reports and benchmarking can be accessed and explored by anyone
- **Speed**:
  - Quicker development feedback; easy to set up for new datasets; enables concurrent work
  - Unlock possibility of adding new features quicker

Suggested links

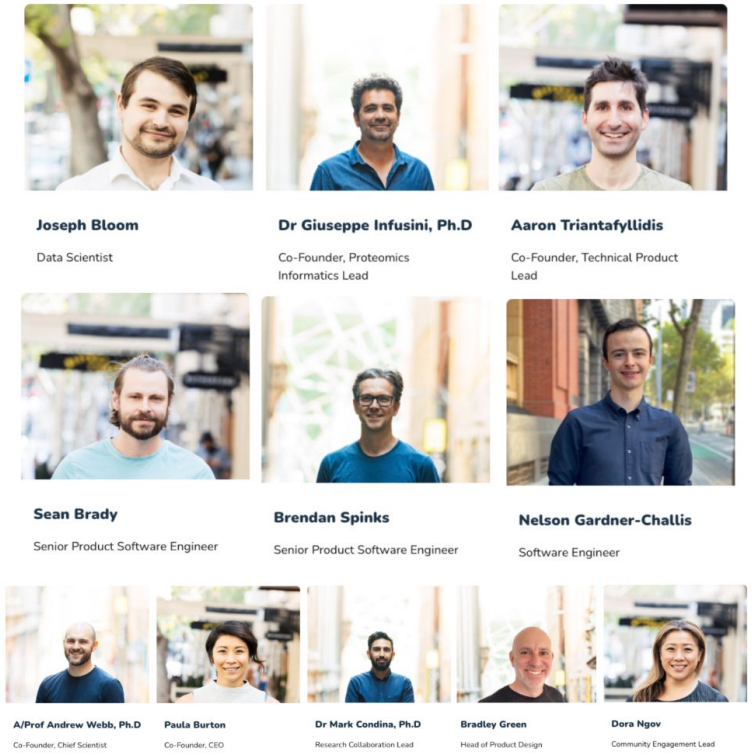
Suggested readings

- Continuous science (2021) Joseph Bloom, Towards Data Science
- DevOps and the scientific process: A perfect Pairing (2022) Christina Hupy, GitLab
- Benchmarking comes of age, (2019), Mark D. Robinson & Olga Vitek, *Genome Biology*

Examples from other fields

- Public benchmarks of robust reading methods: Robust Reading Competition
- From the AI community: Hugging Face

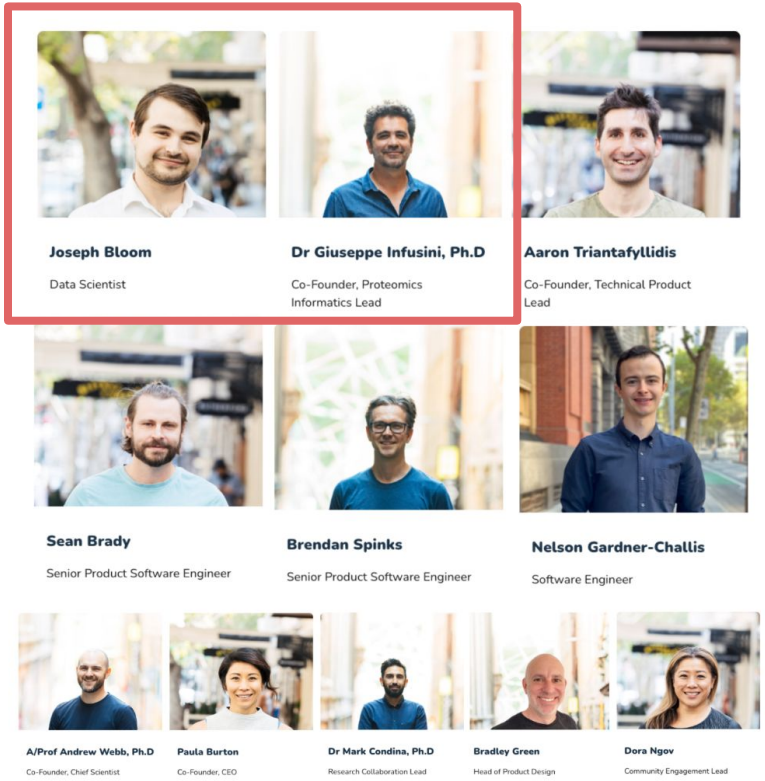
Acknowledgements



Thanks to the Mass Dynamics Team

Booth 727

Acknowledgements

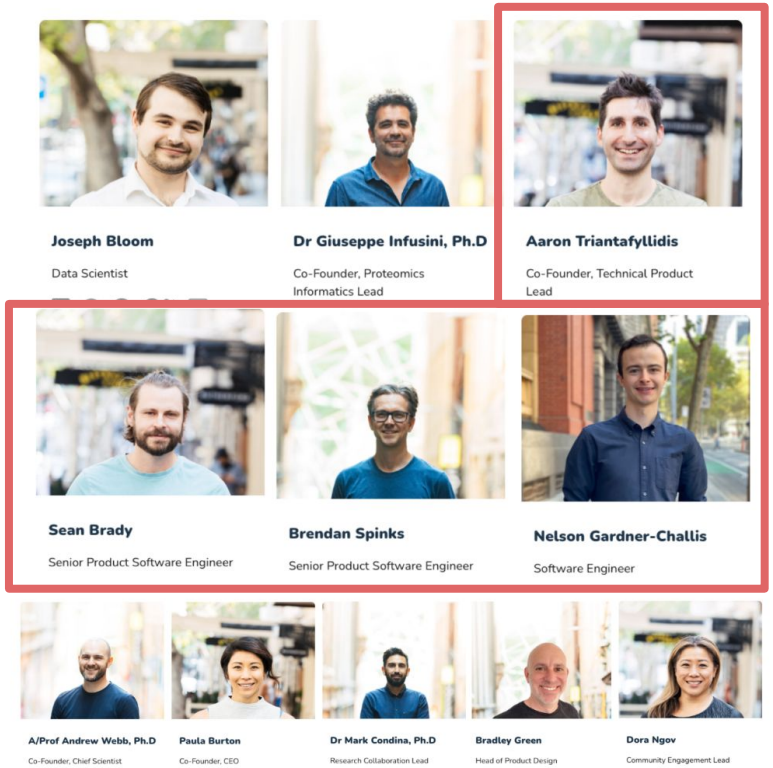


Science team

Thanks to the Mass Dynamics Team

Booth 727

Acknowledgements



Developers team

Thanks to the Mass Dynamics Team

Booth 727

Acknowledgements

PXD009441

Probing the Sensitivity of the Orbitrap Lumos Mass Spectrometer Using a Standard Reference Protein in a Complex Background

Michaella J Levy [1], Michael P Washburn [1] [2], Laurence Florens [1]

Thanks to all the scientists who created the benchmarking data

RETURN TO ISSUE | < PREV **ARTICLE** NEXT >

ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC–MS/MS Experiments

Meena Choi[#†], Zeynep F. Eren-Dogu[#‡], Christopher Colangelo[§], John Cottrell[ǁ], Michael R. Hoopmann[⊥], Eugene A. Kapp[¶], Sangtae Kim[⊗], Henry Lam[□], Thomas A. Neubert[■], Magnus Palmblad[○], Brett S. Phinney[●], Susan T. Weintraub[△], Brendan MacLean[▲], and Olga Vitek[*†]