



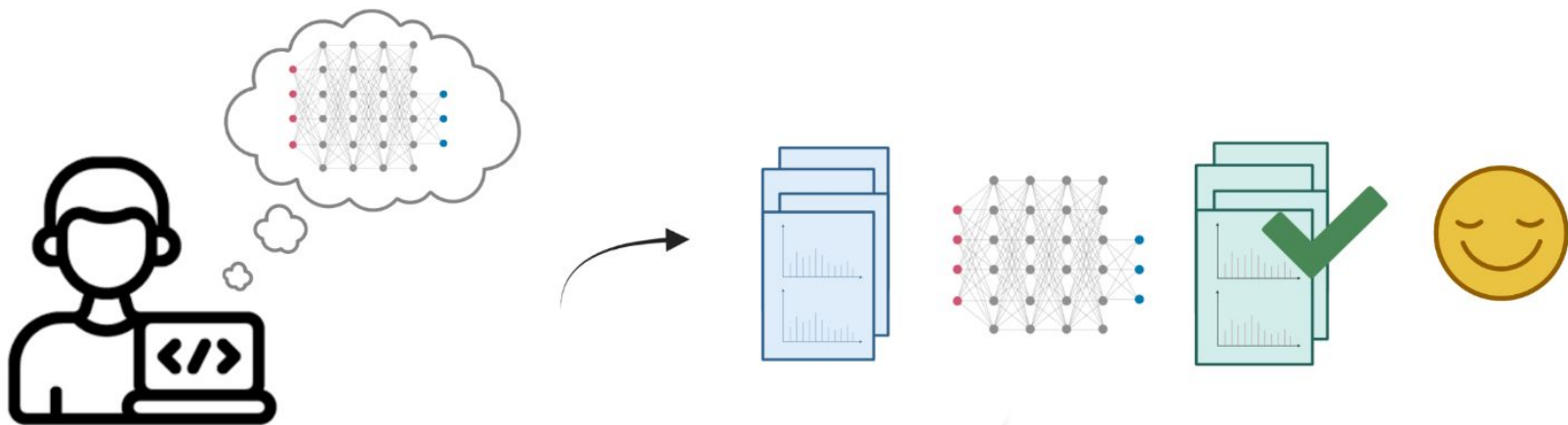
# Automated benchmarking enables continuous confidence in scientific software development

Anna Quaglieri, PhD  
Data Scientist at Mass Dynamics

ASMS Minneapolis  
7th June 2022





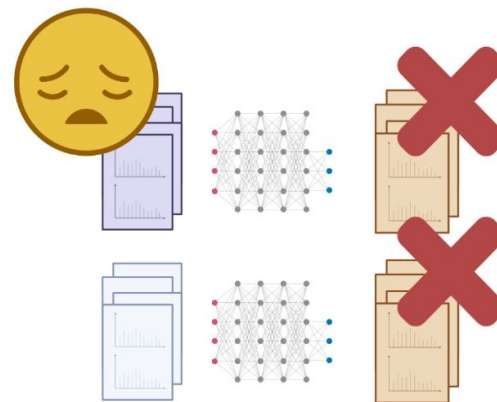






.....

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](#)

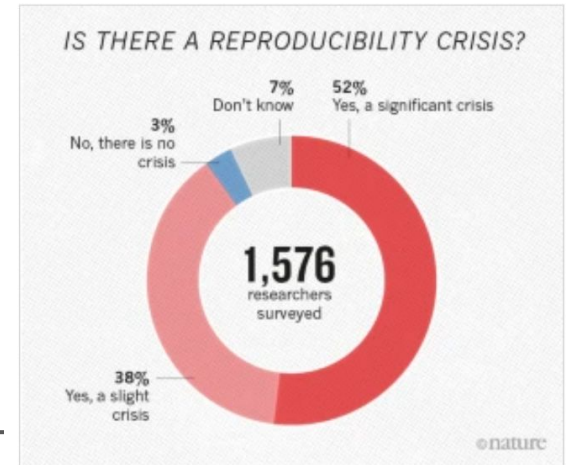
## A manifesto for reproducible science

[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](#)

**222k** Accesses | **1173** Citations | **2593** Altmetric | [Metrics](#)

*“... 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](#)

**44k** Accesses | **1576** Citations | **4014** Altmetric | [Metrics](#)

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

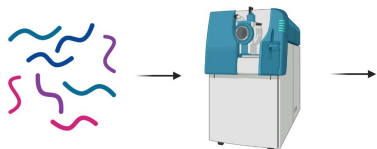
# What I would like you to get from my talk

- 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

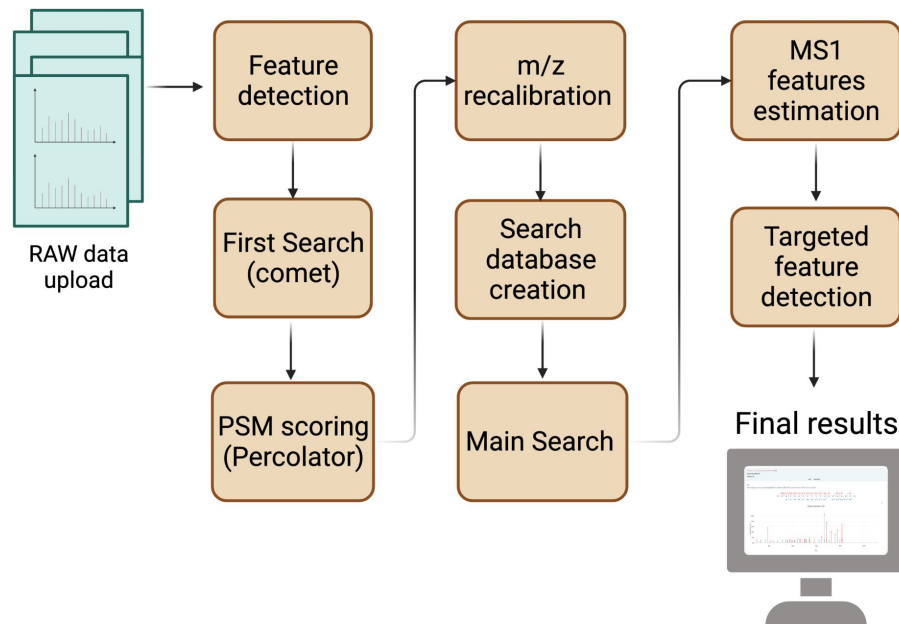
The problem in the proteomics context

# 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***

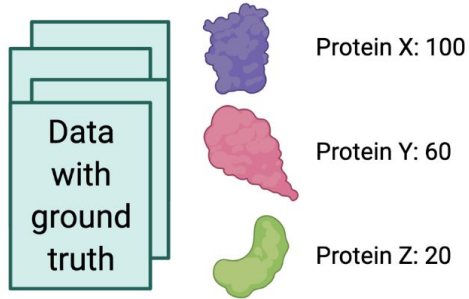
Our solution:  
Automated Benchmarking System

# Our solution: **Automated Benchmarking System**

Brings the CI/CD framework from software into scientific development

Produces automated reports to increase visibility of intermediate steps

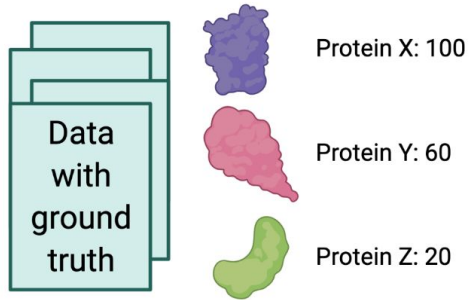
# Benchmarking setup



1 Data with known  
amounts of proteins

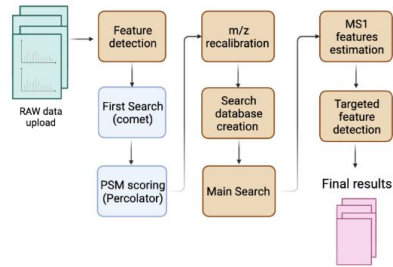
Public and private  
datasets

# Benchmarking setup



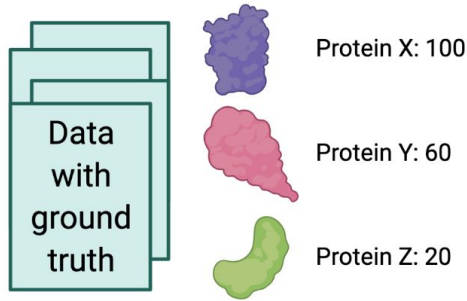
① Data with known amounts of proteins

Public and private datasets



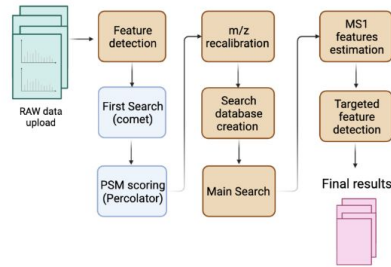
② Workflow to run

# Benchmarking setup



① Data with known amounts of proteins

Public and private datasets



② Workflow to run

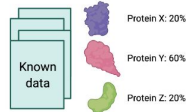


③ Code to check results



# Automated benchmarking system

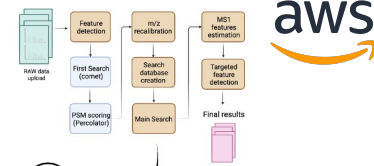
Data with ground truth



1



Workflow



2



3

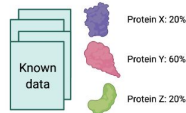


Centralised checks



# Automated benchmarking system

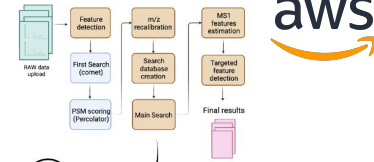
Data with ground truth



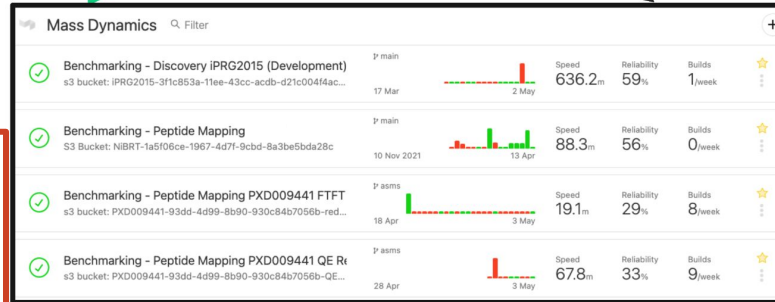
1



Workflow

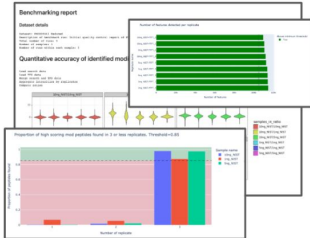


2



Automated tests and reports

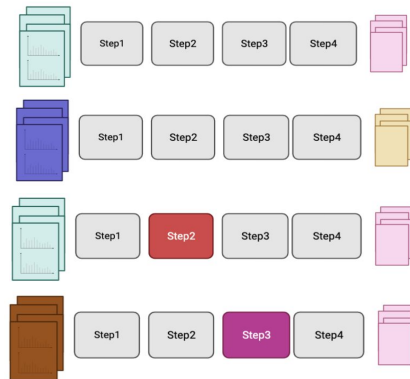
4



Centralised checks

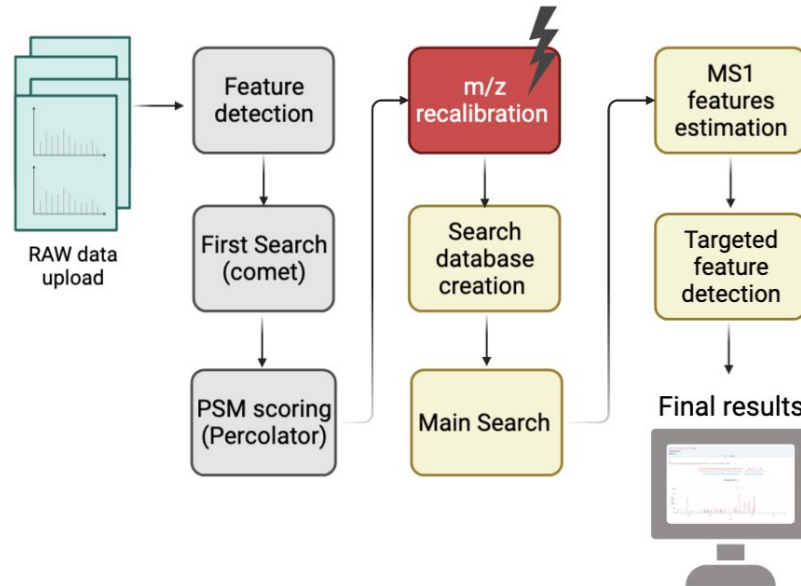


Each **dataset + workflow** has a separate pipeline



# Automated benchmarking in action

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



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

The screenshot displays the Buildkite web interface for a pipeline named "Mass Dynamics / Benchmarking - Peptide Mapping PXD009441". The interface includes a top navigation bar with buttons for "All Builds", "Edit Steps", "Pipeline Settings", and a highlighted "New Build" button (indicated by a red box). Below the navigation bar, the "Test New Workflow" section shows the current build status as "Passed in 10s and blocked". A modal window titled "New Build" is open in the center, allowing users to create a new build. The modal contains fields for "Message" (set to "Test New Workflow"), "Commit" (set to "HEAD"), and "Branch" (set to "main"). A green "Create Build" button is at the bottom of the modal. The background interface also shows a list of build steps, including "buildkite-agent pi...", "Remove deletable tags", and "Prepare python environ...", along with a "Rebuild" button and a "Cancel" button.

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

Test New Workflow  
Build #98 | main | aa47550

buildkite-agent pi...  
Remove deletable tags  
Print

Anna Quaglieri  
Created today at 10:25 | Triggered

✓ buildkite-agent pipeline upl

Passed in 10s and blocked

Prepare python environ...  
Rep...

Rebuild Cancel

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



 **Benchmark**

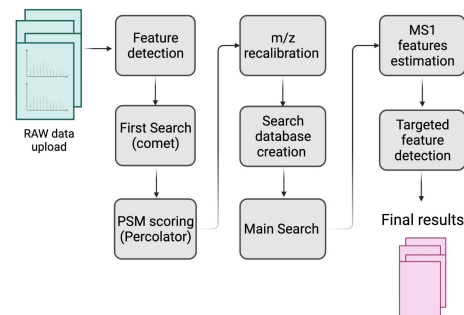
**Do you want to re-run the workflow? — Required**

☒ **Yes, I want to produce new data**


☐ **I will use the data from a previous run**

Continue

## Run the Peptide Mapping Workflow end to end



~ 2-7h




## Benchmark

Do you want to re-run the workflow? — Required

☒ Yes, I want to produce new data  
☐ I will use the data from a previous run

Continue



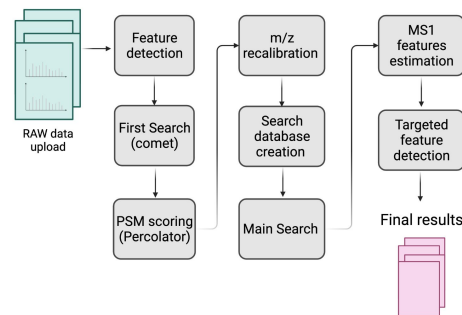
## Benchmark

Do you want to re-run the workflow? — Required

☐ Yes, I want to produce new data  
☒ I will use the data from a previous run

Continue

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 Quagliari

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



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

## Initial benchmarking

Build #3 | main | d281f4d

Failed in 3m 41s



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



Anna Quaglieri

Created Mon 18th Apr at 10:01

Triggered from Web

Rebuild

# Public benchmarking report

app.massdynamics.com/content/workflow/reports

Dataset with ground truth: PXD009441

Levy MG et al, 2018, JPR

## Workflow Benchmarking

Mass Dynamics regularly benchmarks all our workflows, to assess quality and measure performance as we continuously improve.

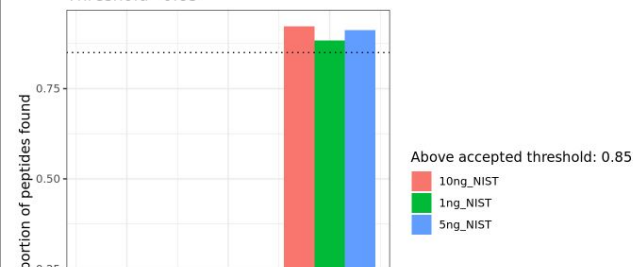
### Peptide Mapping - Probing the Sensitivity of the Lumos Mass Spectrometer using a Standard Reference Protein in a Complex Background\*

The workflow is one of two product characterisation workflows. It is a targeted LFQ-DDA (Label Free Quantification-Data Dependent Acquisition) pipeline involving, feature detection, search, psm-scoring, ms1 recalibration and targeted feature detection (like match-between runs).

\*"Probing the Sensitivity of the Lumos Mass Spectrometer using a Standard Reference Protein in a Complex Background"

Read

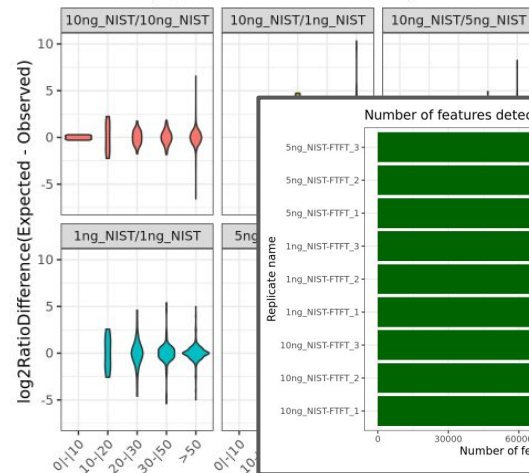
Proportion of modified peptides found in 3or less replicates.  
Threshold=0.85



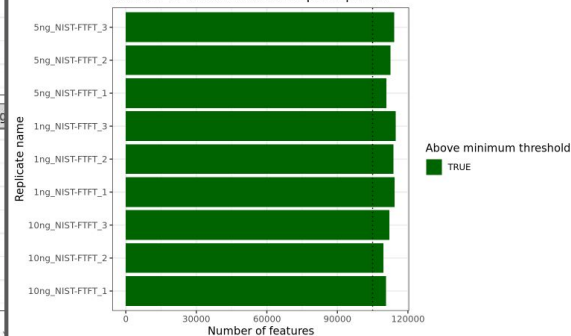
## 3. Quantification accuracy

► How the quantification accuracy is assessed

Modified peptides found in discovery



Number of features detected per replicate



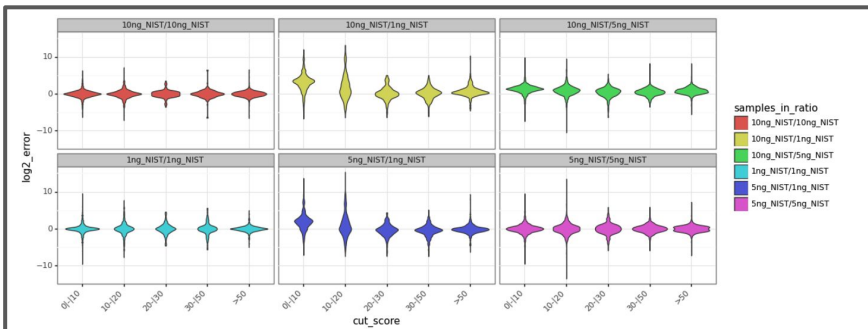
Search score ranges

# How we use the benchmarking system

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

```
1 name: PXD009441
2 description-benchmark: "Initial quality control report of PXD900441"
3 tests:
4   data:
5     number-of-runs: 24
6     number-of-samples: 8
7     replicates-per-sample: 3
8     feature-detection-thermo:
9       minimum-features-per-run: 100000
10    protein-inference:
11      expected-target-above-threshold: 0.001
12      expected-decoy-below-threshold: 0.01
13      q-value-threshold-psms: 0.01
14      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



**Joseph Bloom**

Data Scientist



**Dr Giuseppe Infusini, Ph.D**

Co-Founder, Proteomics  
Informatics Lead



**Aaron Triantafyllidis**

Co-Founder, Technical Product  
Lead



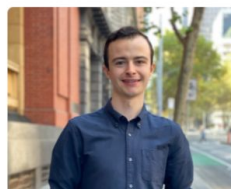
**Sean Brady**

Senior Product Software Engineer



**Brendan Spinks**

Senior Product Software Engineer



**Nelson Gardner-Challis**

Software Engineer



**A/Prof Andrew Webb, Ph.D**

Co-Founder, Chief Scientist



**Paula Burton**

Co-Founder, CEO



**Dr Mark Condina, Ph.D**

Research Collaboration Lead



**Bradley Green**

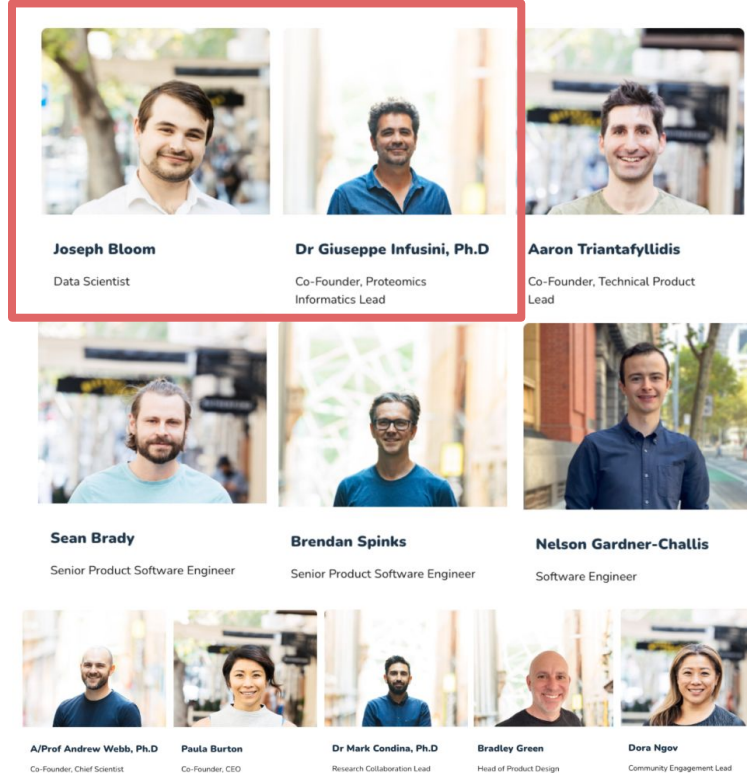
Head of Product Design



**Dora Ngov**

Community Engagement Lead

# Acknowledgements



Thanks to the  
Mass Dynamics Team

Booth 727

Science team

# Acknowledgements



**Joseph Bloom**

Data Scientist



**Dr Giuseppe Infusini, Ph.D**

Co-Founder, Proteomics  
Informatics Lead



**Aaron Triantafyllidis**

Co-Founder, Technical Product  
Lead



**Sean Brady**

Senior Product Software Engineer



**Brendan Spinks**

Senior Product Software Engineer



**Nelson Gardner-Challis**

Software Engineer



**A/Prof Andrew Webb, Ph.D**

Co-Founder, Chief Scientist



**Paula Burton**

Co-Founder, CEO



**Dr Mark Condina, Ph.D**

Research Collaboration Lead



**Bradley Green**

Head of Product Design



**Dora Ngov**

Community Engagement Lead

## Thanks to the Mass Dynamics Team

Booth 727

## Developers team

# Acknowledgements

PXD009441

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

Michaela J Levy<sup>1</sup>, Michael P Washburn<sup>1 2</sup>, Laurence Florens<sup>1</sup>

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<sup>#†</sup>, Zeynep F. Eren-Dogu<sup>#‡</sup>, Christopher Colangelo<sup>§</sup>, John Cottrell<sup>‡</sup>, Michael R. Hoopmann<sup>‡</sup>, Eugene A. Kapp<sup>‡</sup>, Sangtae Kim<sup>®</sup>, Henry Lam<sup>□</sup>, Thomas A. Neubert<sup>■</sup>, Magnus Palmblad<sup>°</sup>, Brett S. Phinney<sup>•</sup>, Susan T. Weintraub<sup>△</sup>, Brendan MacLean<sup>▲</sup>, and Olga Vitek<sup>\*\*</sup>