Case study: publishing an interactive, reproducible workflow using Frictionless data standards

Varvara Efremova [a], James Wilmot [a], Prof. Pall Thordarson [a,b]

[a] UNSW School of Chemistry, The Australian Centre for Nanomedicine, [b] The UNSW RNA Institute

opendatakit

Closing the gap between analysis and publication

To achieve fully FAIR web publication of interactive workflows, we developed a set of tools alongside a workflow packaging specification.

Our tooling allows users to run and publish fully FAIR workflows as embeddable web components.

Datakits & Frictionless Data

In order to encapsulate analysis workflows, we developed "datakits" - an open-source JSON-based data standard extending the Frictionless Data specification. (See `frictionlessdata.io`)

A datakit describes:

- the analysis algorithm and its execution environment
- saved run states from algorithm executions
- input and output data, along with configurable options
- visualisations of data, including graph and table specifications
- user interface definitions.

Publishing a datakit - binding constant analysis

Fork

Fork the `open-datakit/bindfit-datakit` repository to preserve provenance

Ingest data

1. Load data and automatically describe schema
   `dk load data data/input.csv`

Configure algorithm

1. Choose fit model
   `dk set model nmr1to2`
2. Set initial parameter guesses
   `dk set params k11.init 1.44`

Run algorithm

`dk run`

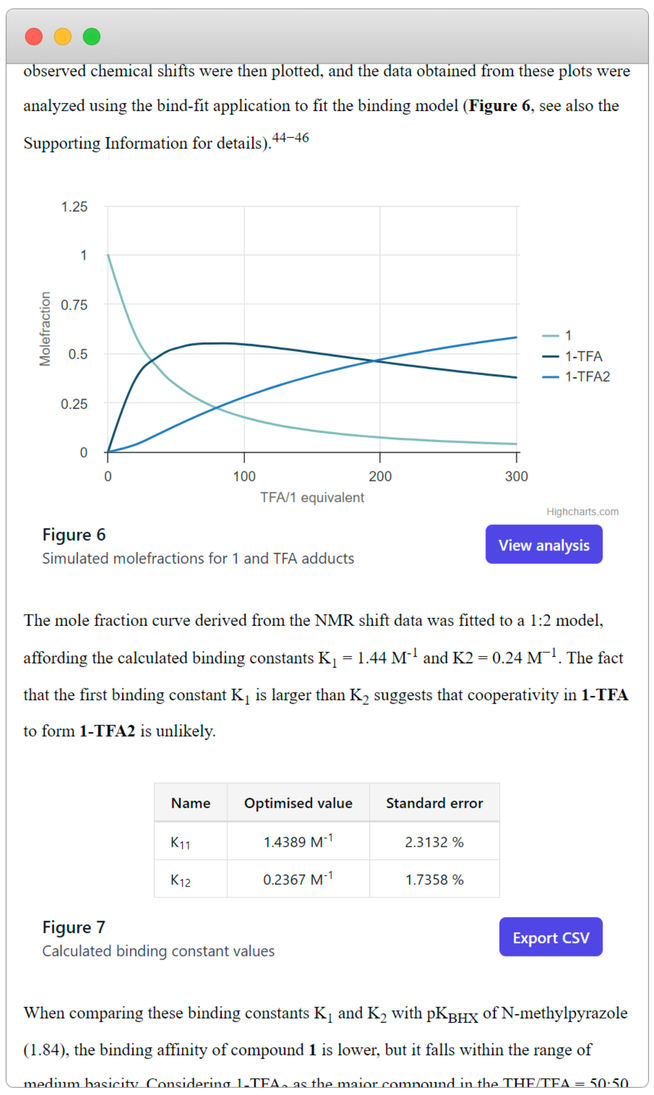
Visualise

Visualise facets of the analysis

1. View fit graph
   `dk view fitGraph`
2. View calculated parameter table
   `dk show params`

Publish

1. Create a release of your datakit workflow and mint a DOI with zenodo



observed chemical shifts were then plotted, and the data obtained from these plots were analyzed using the bind-fit application to fit the binding model (**Figure 6**, see also the Supporting Information for details).[44–46]

Figure 6
Simulated molefractions for 1 and TFA adducts

View analysis

The mole fraction curve derived from the NMR shift data was fitted to a 1:2 model, affording the calculated binding constants K_1 = 1.44 M^{-1} and K2 = 0.24 M^{-1}. The fact that the first binding constant K_1 is larger than K_2 suggests that cooperativity in **1-TFA** to form **1-TFA2** is unlikely.

Name	Optimised value	Standard error
K_{11}	1.4389 M^{-1}	2.3132 %
K_{12}	0.2367 M^{-1}	1.7358 %

Figure 7
Calculated binding constant values

Export CSV

When comparing these binding constants K_1 and K_2 with pK_{BHX} of N-methylpyrazole (1.84), the binding affinity of compound **1** is lower, but it falls within the range of

Figure 1: Embeddable workflows in an example publication

Technical details

We focused on a low-code approach, making use of existing industry-standard tools whenever possible:

- Frictionless Data specification
- Docker containerisation for execution environments
- Git repositories for history tracking
- GitHub for analysis publication.

Our command line tool is written in Python. Embeddable web components are implemented in SolidJS with solid-element, making them usable on any website.

Proof of concept: Bindfit

In 2016, we released Bindfit — a web application for interactive analysis and publication of binding constant calculations. Every Bindfit workflow can be saved and published via a link with a unique identifier.

Since 2016, it has gained wide adoption with over 70,000 sessions and nearly 500 citations on Scopus.

Example of a published analysis in Bindfit
(Takemasa & Nozaki 2024)
`app.supramolecular.org/bindfit/view/b24c676b-5fc5-4774-8caa-9dc1267ee771`

The opendatakit project

Currently released:

- a binding constant analysis tool packaged as a datakit
- a command line tool for modifying and running datakits
- documentation for running and implementing custom datakits.

In the pipeline:

- releasing embeddable web components for datakit visualisations
- developing a web application with execution backend
- handling domain-specific metadata
  (Please come to our BoF! *"Data and metadata standards for web publication of research data"*)
- providing consultancy on datakit construction and visualisations.

If you're interested in building datakits or collaborating, please get in touch! **hello@opendatakit.io**

Further reference

Datakit and command line documentation
`docs.opendatakit.io`

Command line tool
`open-datakit/cli`

Bindfit datakit
`open-datakit/bindfit-datakit`

UNSW RNA Institute

UNSW SYDNEY