First steps to data portals: Globus, data catalog, and You

Andy Boughton
(abought@globus.org)
APECx Demos – March 2025

Globus Transfer: Part of an integrated platform

Learn more:

https://www.globus.org/platform

https://www.go-fair.org/fair-principles/

Today's notes:

https://github.com/globus/apecx-demos/

Try it yourself

Get an endpoint: https://www.globus.org/globus-connect-personal

Manage files: https://app.globus.org/

Example static search portal: https://abought.github.io/apecx-demo-static-search-portal

-> Real version *not yet publicly available*

Findable

Accessible

Interoperable

Reusable

https://www.go-fair.org/fair-principles/
https://www.youtube.com/watch?v=N2zK3sAtr-4

SPHERICAL About

andrew.boughton@globus.org - Logout

SPHERICAL: Antigens by Design

An APECx Project

The SPHERICAL data repository contains inputs from many teams. Explore example datasets below:

Measure

Experimental data

Protein structure examples

Predict

Computational models

Demo render files

Insights

Curated insights from across the community

Publications

Literature chatbot

*

Type

- image (5)
- large (5)
- text (5)
- table (4)
- code (3)
- 404 (2)
- csv (2)
- empty (2)
- fail (2)
- tsv (2)
- embed (1)
- gzip (1)
- javascript (1)
- jpg (1)
- json (1)
- pdf (1)
- png (1)
- protein (1)
- python (1)
- resize (1)
- svg (1)

Results

19 datasets found

Publication Type

- JOUR (24)
- CHAP (1)
- COMP (1)
- CONF (1)
- THES (1)

Top Tags

- chemistry (14)
- spectroscopy (14)
- genetics (10)
- protein (8)
- software (8)
- visualization (5)
- crowdsourcing (4)
- cell imaging (3)
- covid (3)
- machine learning (3)

Big image

Description: A big background picture

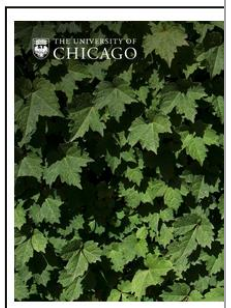


image , resize

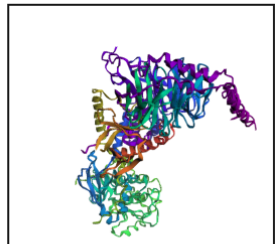
JSON blob

Description: A json blob with odd formatting, to test code highlighting of rendered JSON

{"key":"value", "items": [1,2,3], "nothing": null, "nested": {}}

1OMW

DOWNLOAD VIA HTTPS | RUN CALCULATION



PDB ID: 1OMW

DOI: https://doi.org/10.2210/pdb1OMW/pdb
Description: Crystal Structure of the complex between G Protein-Coupled Receptor Kinase 2 and Heterotrimeric G Protein beta 1 and gamma 2 subunits

Organism: bos taurus

Deposited: 2003-02-26T00:00:00+00:00
Released: 2003-06-03T00:00:00+00:00
Authors: Lodowski, D.T.; Pitcher, J.A.; Capel, W.D.; Lefkowitz, R.J.; Tesmer, J.J.G.
Funders: (none provided)

protein , gprotein , gpcr , xrd , experimental

Demo of portal-in-progress

(not public yet)

Core Technologies

Technology highlights

- **Globus provides:**
  - One auth system across services
  - Groups to control access
  - Full text search features
- **Reusable tools**
  - Python and JavaScript SDKs
  - “Search portal” helper tools
  - Prebuilt UI for workflows and file transfer

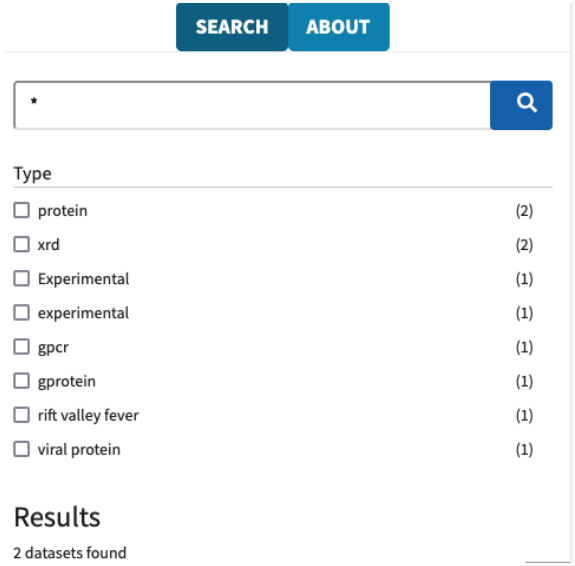
Django Globus Portal Framework: Build search-focused data portals https://github.com/globus/django-globus-portal-framework
Python social auth: Generic auth toolkit for Globus and python webapps https://github.com/python-social-auth/social-core/blob/master/social_core/backends/globus.py

README Apache-2.0 license

Django Globus Portal Framework

DOI 10.5281/zenodo.14919979 Tests passing pypi v0.4.13 wheel yes License Apache 2.0

The Django Globus Portal Framework is a collection of tools that enable you to rapidly create an easily accessible portal for your data. Globus provides robust Auth and Search services, both powerful tools to help manage who has access to your data and metadata. Tailoring your portal to your data can be done simply by modifying existing bulit-in Bootstrap templates, allowing many levels of customization to suit the required needs.



`settings.py` file to enable user auth and Globus Portal Framework components. You can copy-paste the individual settings below, or use our Settings Example for a complete `settings.py` file reference.

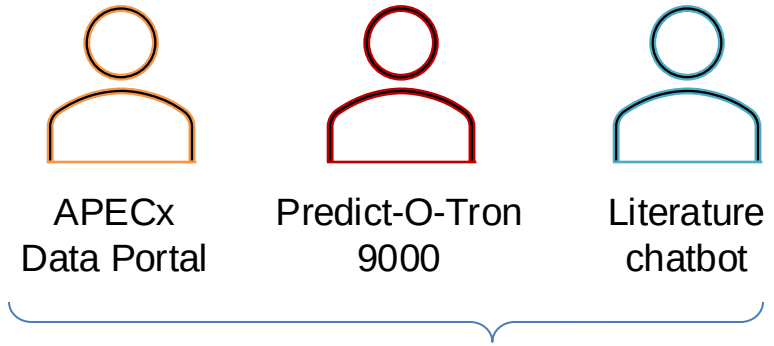
```
# Your portal credentials for a Globus Auth Flow
SOCIAL_AUTH_GLOBUS_KEY = 'Put your Client ID here'
SOCIAL_AUTH_GLOBUS_SECRET = 'Put your Client Secret Here'

# This is a general Django setting if views need to redirect to login
# https://docs.djangoproject.com/en/3.2/ref/settings/#login-url
LOGIN_URL = '/login/globus'

# This dictates which scopes will be requested on each user login
SOCIAL_AUTH_GLOBUS_SCOPE = [
    'urn:globus:auth:scope:search.api.globus.org:search',
]

# Installed apps tells Django which packages to load on startup
INSTALLED_APPS = [
    ...
    'globus_portal_framework',
    'social_django',
]
```

Build your own adventure

- **SDKs control every aspect of Globus**
  - Python: https://globus-sdk-python.readthedocs.io/en/stable/
  - JS: https://github.com/globus/globus-sdk-javascript
- **Higher-order helpers build websites fast**
  - Django: https://github.com/globus/django-globus-portal-framework
- **Reusable libraries let you adopt just one feature**
  - Python Social auth: https://github.com/python-social-auth/social-core/blob/master/social_core/backends/globus.py

Findable and reusable

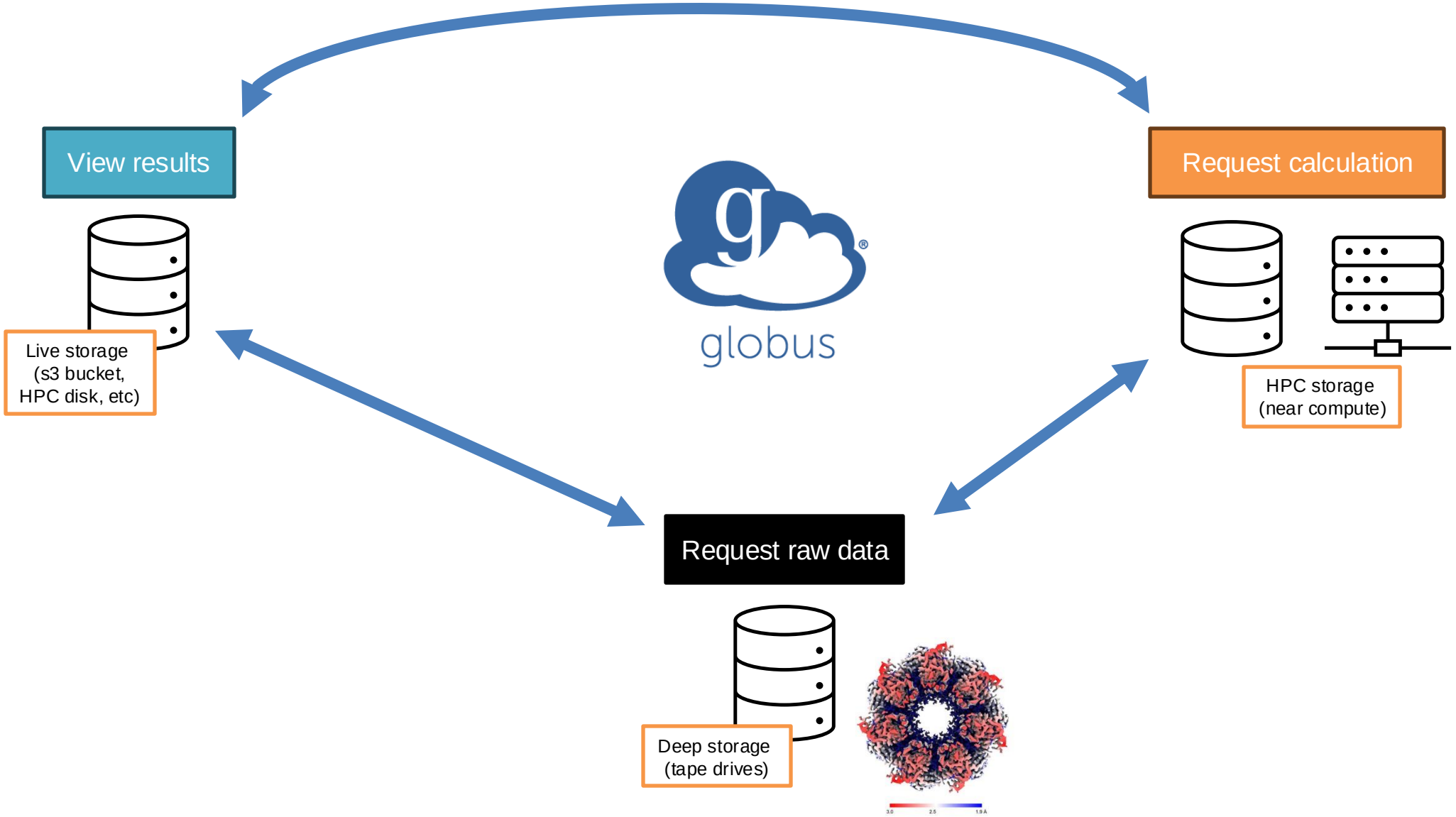
- **Multiple websites can build on the shared data catalog / portal**
- **Shared foundation allows reuse of data**
- **We do the fiddly stuff (“members of APECx, only”)**



One auth system

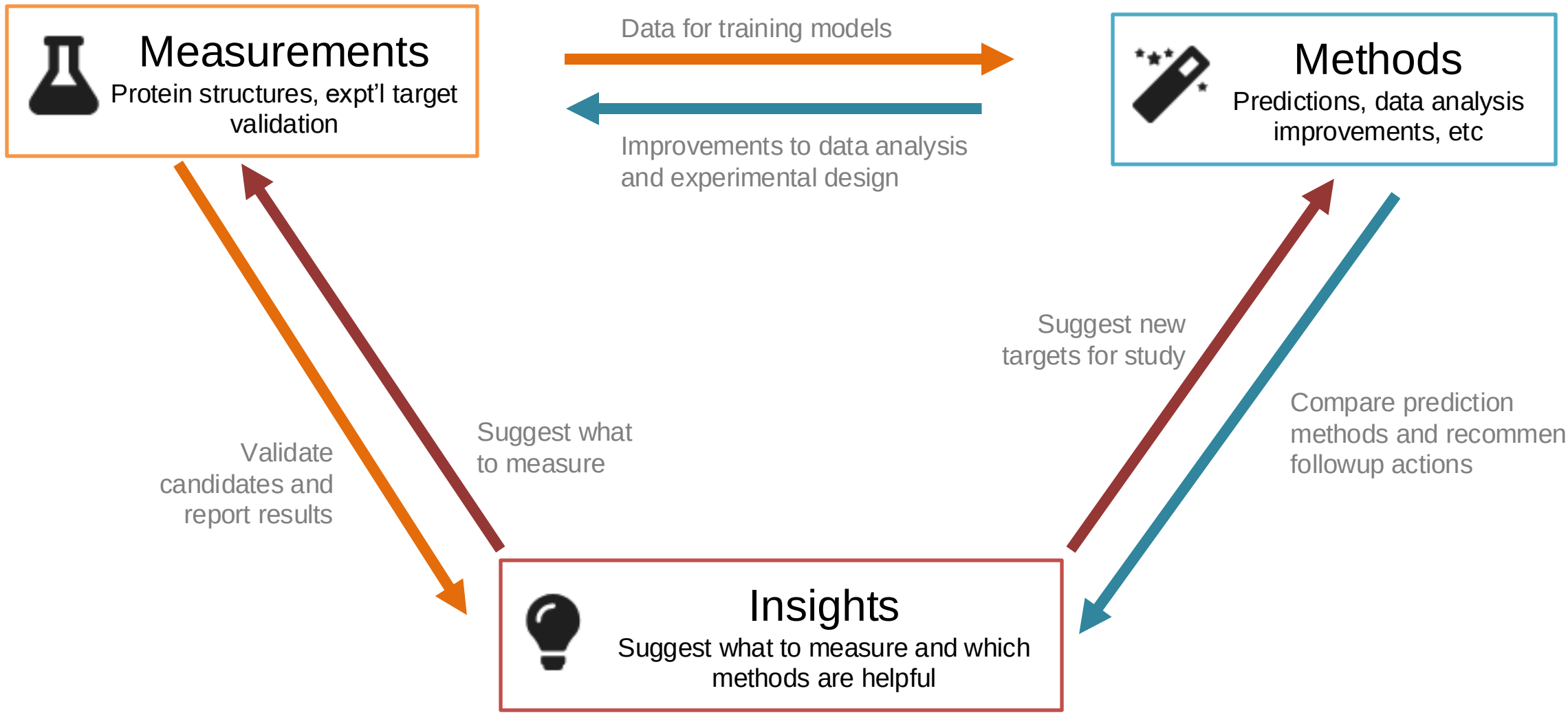
Consistent access rules

Source training data from a shared catalog

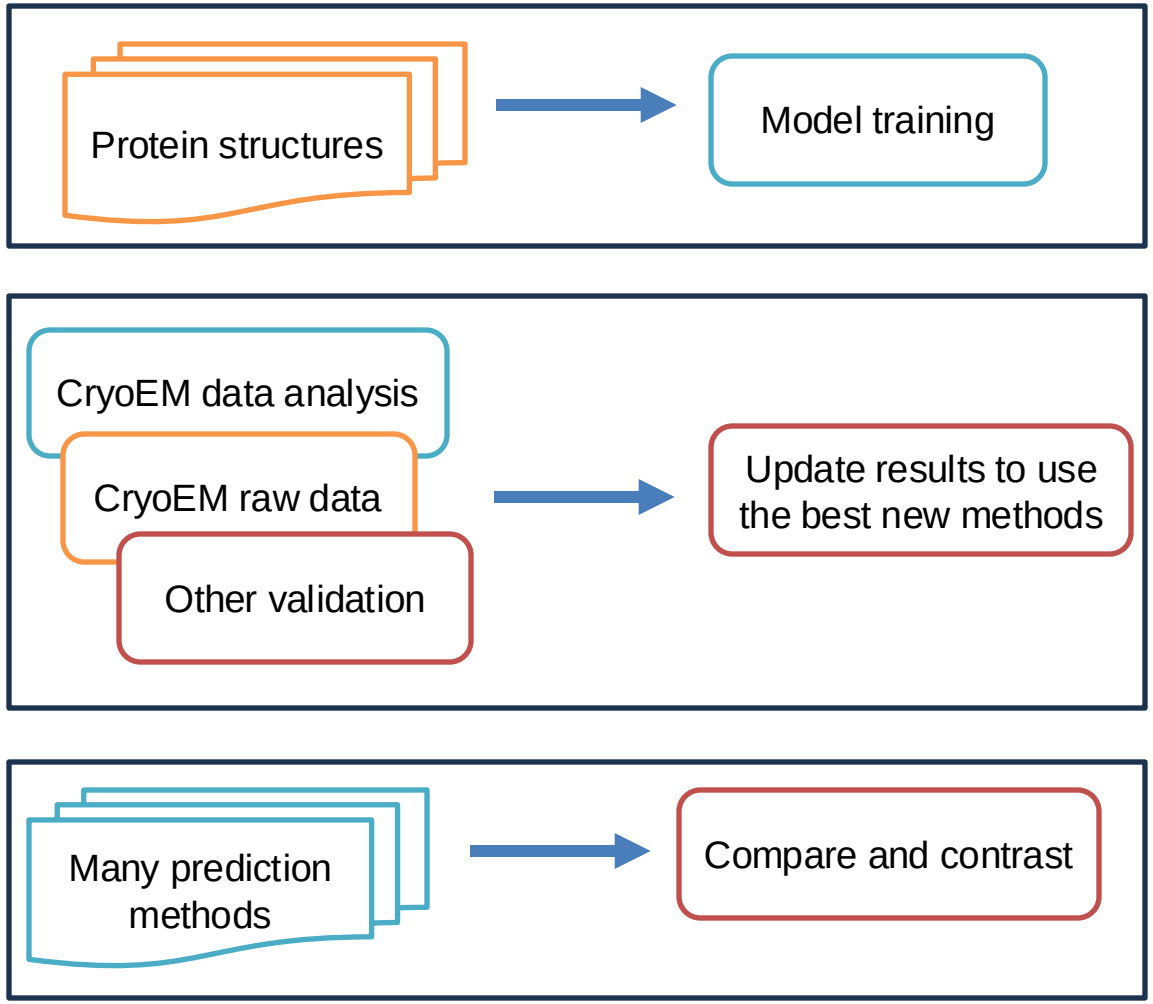

View results
Request calculation
globus
Live storage
(s3 bucket,
HPC disk, etc)
HPC storage
(near compute)
Request raw data
Deep storage
(tape drives)

All the fancy stuff depends on data

Where do we start?


Measurements
Protein structures, expt'l target validation
Data for training models
Improvements to data analysis and experimental design
Methods
Predictions, data analysis improvements, etc
Validate candidates and report results
Suggest what to measure
Suggest new targets for study
Compare prediction methods and recommend followup actions
Insights
Suggest what to measure and which methods are helpful

Points of collaboration and connection



- **Create categories based on value-add capabilities**
  - Identify “AI/ML ready” models
  - Run a method that requires inputs of a certain type
  - Incorporate QC for things that might feed into methods
- **Search features that allow discovering reusable items**

FAIR: Findable, Accessible, Interoperable, Reusable
https://www.go-fair.org/fair-principles/

Finding a Taxonomy

- **Encourage discovery across research groups**
  - Identify what datasets have in common
- **Share a common language**
  - Eg all submissions should use the same molecule/virus name/ID symbol
- **Unrealistic to build custom apps for every individual dataset- what are the big wins?**

Some example data types (maybe)

Protein structures (measured and predicted)

Predicted druggable targets

In silico estimates of antibody efficacy

Clinical validation

Available squishy wet lab things

…Many others

Catalog: APECx models and methods v1 – Jan 2025
https://docs.google.com/spreadsheets/d/1iwR-hpcMyvBXTfmE-Nf9AkJpAfp5LpOaoep3U-Z1mIs/preview

Discussion questions

Audience Questions: Data Catalog

- **What are some patterns in the data types we have catalogued to date?**
- **Is there a standard ID format / language for entities?**
- **What general categories would encourage interoperable re-use?**

Some example data types (maybe)

Protein structures (measured and predicted)

Predicted druggable targets

In silico estimates of antibody efficacy

Clinical validation

Available squishy wet lab things

…Many others

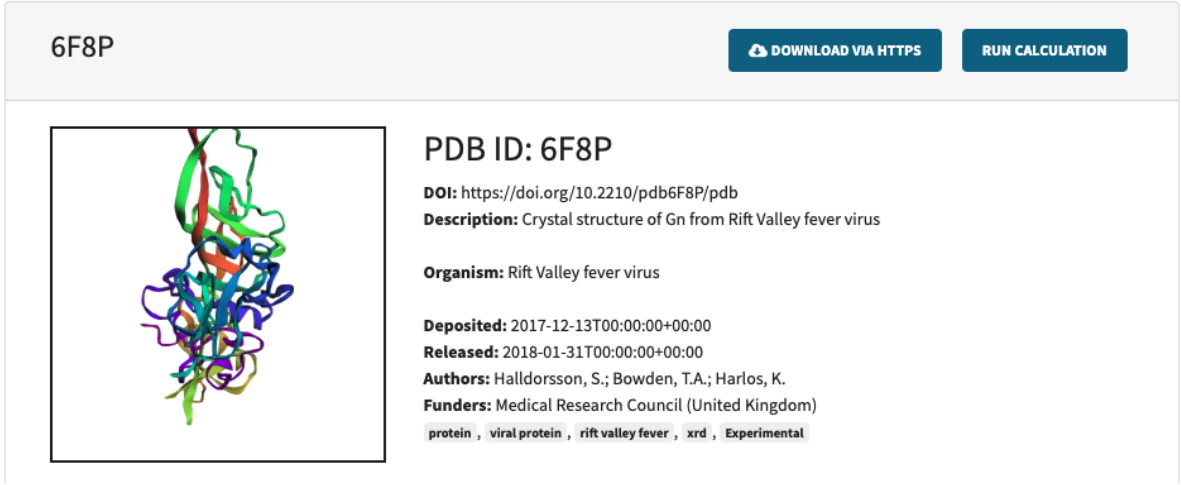
Measurements
Protein structures, expt'l target validation

Methods
Predictions, data analysis improvements, etc

Insights
Suggest what to measure and which methods are helpful

Audience Questions: First goal

- **What is a highly reusable piece of data that multiple groups want to access?**
  - And what is needed to make that data “AI/ML ready”
- **What is a useful presentation of that data?**
  - How does a user decide which dataset will meet their needs?
  - Schema + interactive features



Audience questions: General

- **What topics would be useful for a future talk?**
  - Globus technologies (workflows/ compute, auth, etc)
  - Deep dive on data
  - Portal features
- **What do we want the portal to *do*?**
  - We have the technology
  - You have the science

Proposed “user stories”: https://docs.google.com/document/d/10bCWtOQXZd8fJS3nPCYwlCcYozgraTELM1UfNZBGzwk/preview

Future questions

- **Where does the website live?**
- **Who has access?**