SBGrid Databank

Peter A Meyer[1], Stephanie Socias[1], Jason Key[1],
Mercè Crosas[2], Piotr Sliz[1,3]

1 BCMP, Harvard Medical School and SBGrid Consortium;
2 IQSS, Harvard University and Dataverse Project;
3 Boston Children's Hospital, Dept. of Pediatrics

https://data.sbgrid.org

Problem

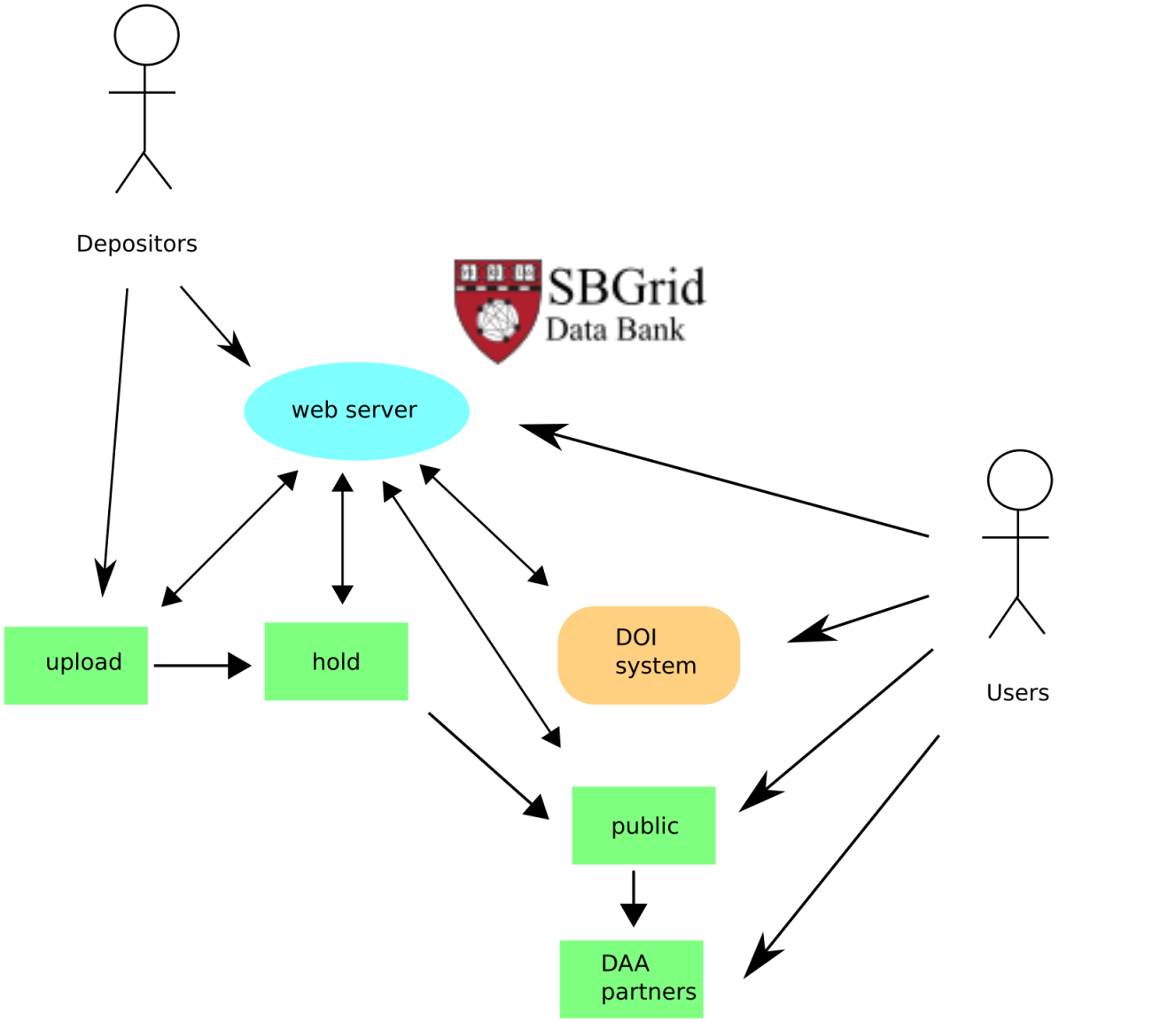
- Diffraction images for published structures not widely available
  - Limits re-analysis and validation
  - Difficult to use as controls
  - Limits method development
  - May impact funder data retention policies

Why haven’t images been as widely available as PDBs?

- Needs more storage [10s GB vs 10s MB]
- Wider format variations
- Concerns about how reusable images will be
- It’s more work for experimenters

What's a dataset?

- Generically
  - Data files: inputs to analysis/compute
  - Metadata: information needed to use data files
- Primary X-ray diffraction datasets
  - Image files
  - Resulting in single Intensity/Amplitude set



INSTITUT PASTEUR
DE MONTEVIDEO

SIBS
SHANGHAI INSTITUTES FOR BIOLOGICAL SCIENCES,
CHINESE ACADEMY OF SCIENCES

UPPSALA
UNIVERSITET

Getting Data Out

Data Access Instructions

1. If this dataset is locally available, it should be accessable at `/programs/datagrid/1`
2. To download this dataset, please run the following command from your Terminal on a Linux or OS X workstation:

`'rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/1 .'` (Harvard Medical School, USA)

Depending on your location, faster access may be available from a Tier 1 site closer to your location

`'rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/1 .'` (Uppsala University, Sweden)

`'rsync -av rsync://sbgrid.pasteur.edu.uy/10.15785/SBGRID/1 .'` (Institut Pasteur de Montevideo, Uruguay)

`'rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/1 .'` (Shanghai Institutes for Biological Sciences, China)

3. After the transfer is completed, please issue the following command to verify data integrity:

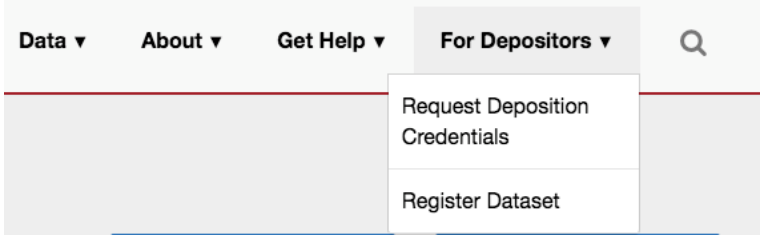
`'cd 1 ; shasum -c files.sha'`

Storage requirements: 1.6G

```
pameyer@pm-linux:~$ rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/1 .
receiving incremental file list
1/
1/files.sha
1/p3_6_1.runlist
1/p3_6_1_001.img
1/p3_6_1_002.img
1/p3_6_1_003.img
1/p3_6_1_004.img
1/p3_6_1_005.img
1/p3_6_1_006.img
1/p3_6_1_007.img
1/p3_6_1_008.img
1/p3_6_1_009.img
1/p3_6_1_010.img
1/p3_6_1_011.img
1/p3_6_1_012.img
1/p3_6_1_013.img
1/p3_6_1_014.img
```

```
pameyer@pm-linux:/programs/datagrid/1$ shasum -c files.sha
./p3_6_1_006.img: OK
./p3_6_1_001.img: OK
./p3_6_1_073.img: OK
./p3_6_1_008.img: OK
./p3_6_1_074.img: OK
./p3_6_1_090.img: OK
./p3_6_1_041.img: OK
./p3_6_1_046.img: OK
./p3_6_1_048.img: OK
./p3_6_1_034.img: OK
./p3_6_1_033.img: OK
./p3_6_1_022.img: OK
./p3_6_1_059.img: OK
./p3_6_1_088.img: OK
./p3_6_1_025.img: OK
./p3_6_1_057.img: OK
```

Putting Data In



Depositor Name:

Depositor Email:

Depositor Institution:

PI Name:

PI Email:

PI Institution:

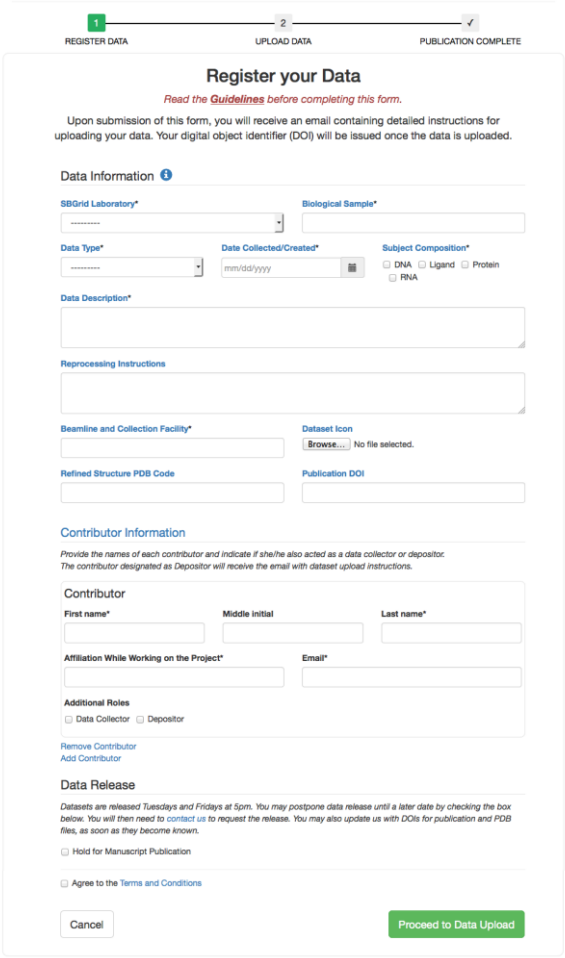
PI ORCID:

Lab website:

This is a request from a new lab or group: ☑

Captcha: CNS

Request Credentials




```
pameyer@pm-linux:~$ bash ./upload2sbdb_dataset-No0.bash
Please enter the full path of the directory containing dataset files to upload for dataset number: 0
This directory should contain only the dataset in question - all files within this directory will be uploaded
/data/images/
files to be uploaded in /data/images/:
/data/images/p3_6_1_001.img
/data/images/p3_6_1_002.img
/data/images/p3_6_1_003.img
/data/images/p3_6_1_004.img
/data/images/p3_6_1_005.img
/data/images/p3_6_1_006.img
/data/images/p3_6_1_007.img
/data/images/p3_6_1_008.img
/data/images/p3_6_1_009.img
/data/images/p3_6_1_010.img
```

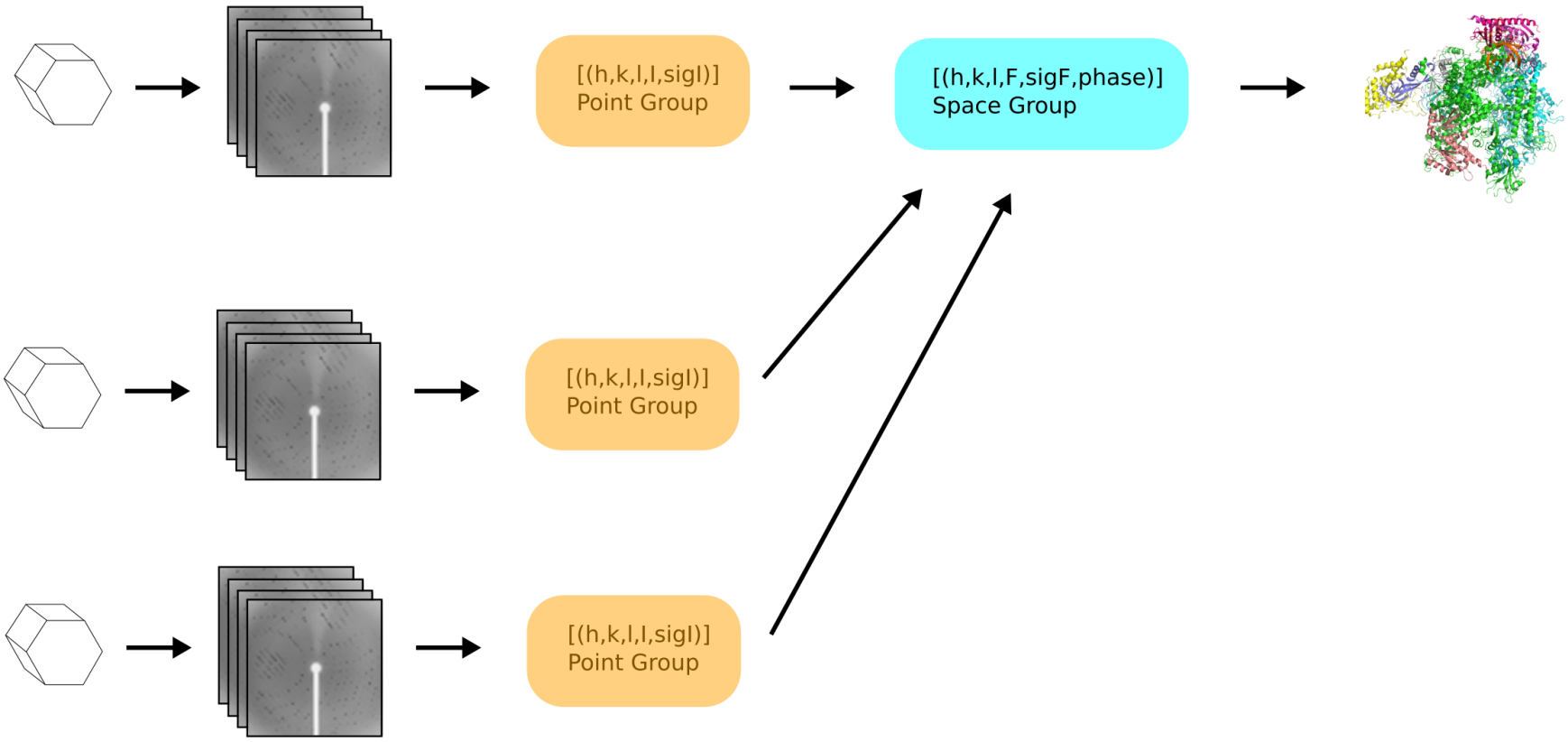
X-Ray Diffraction Collection Changes

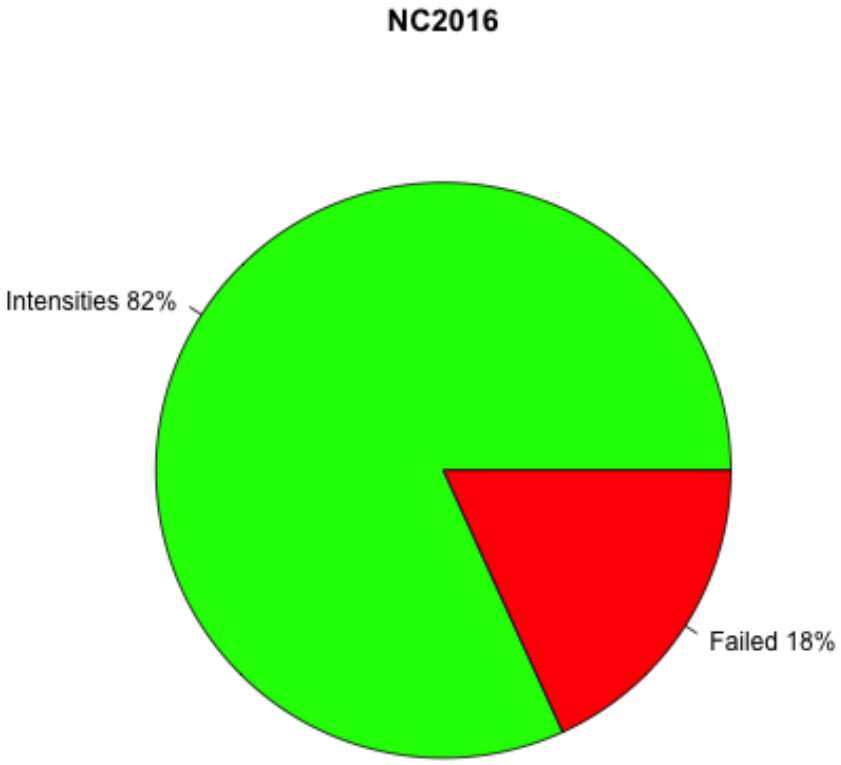
	NC 20150901	20170524
Datasets	111	311
PDBs	91	274
Articles	65	128
Labs / Groups	48	66
Institutions	34	42
Collection Facilities*	65	151
Detectors*	8	23

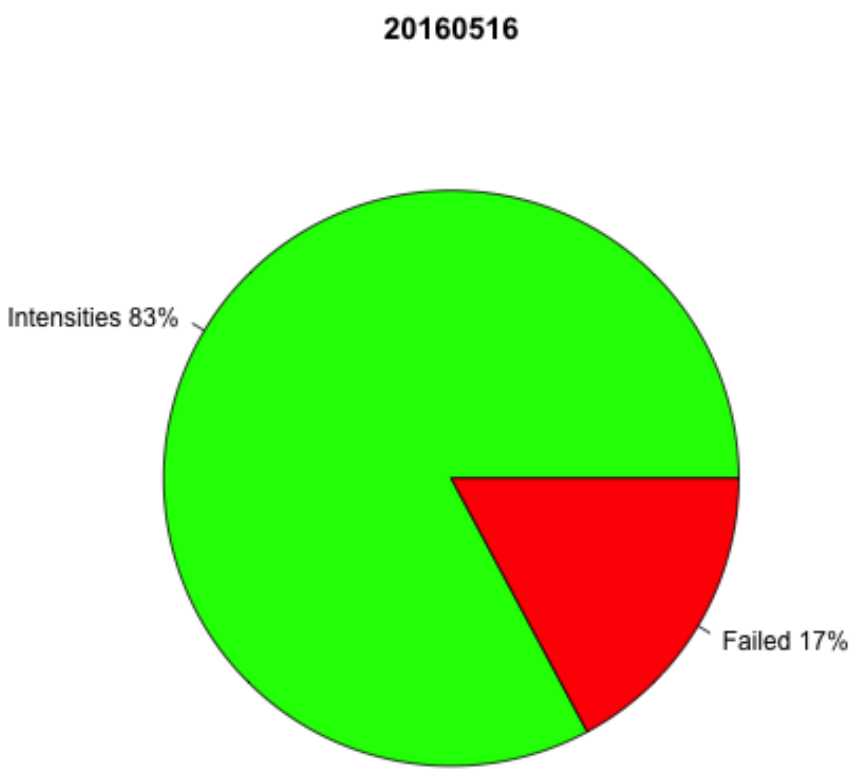
doi:10.1038/ncomms10882

How reusable is this data?

- Crystallographic structure determination is multi-step
- Reusability needs to evaluated stepwise
- With known structures, can be evaluated “backwards”


[(h,k,l,I,sigI)]
Point Group
[(h,k,l,I,sigI)]
Point Group
[(h,k,l,I,sigI)]
Point Group
[(h,k,l,F,sigF,phase)]
Space Group





110 datasets

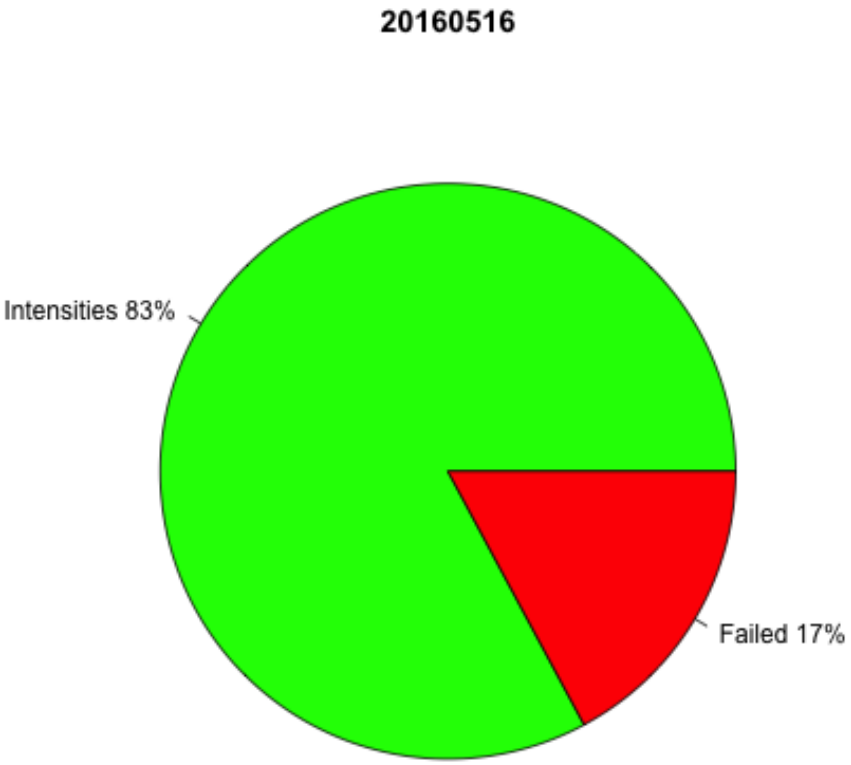
309 datasets

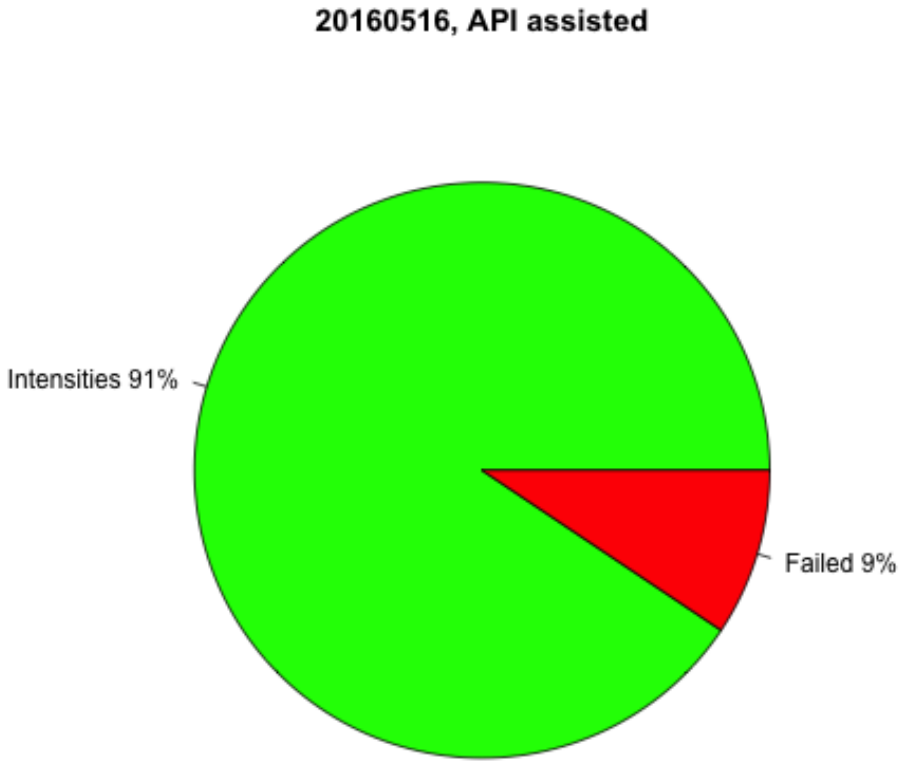
How to get better metadata?

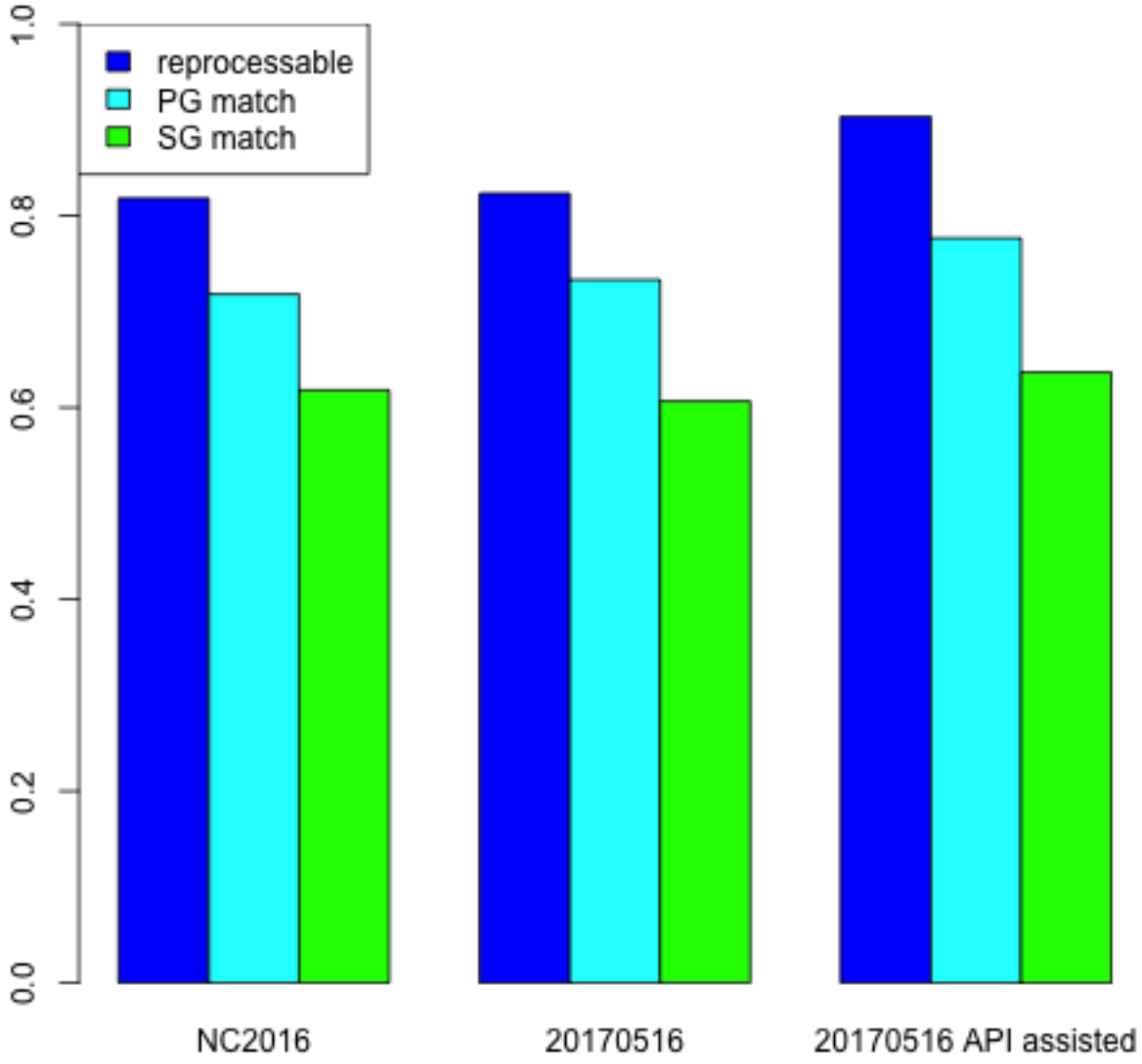
- Encourage and support standardized formats with accurate experimental metadata
- Ask depositors to provide experimental parameters
- Have curators attempt to to determine experimental metadata
- Have computers attempt to determine experimental metadata

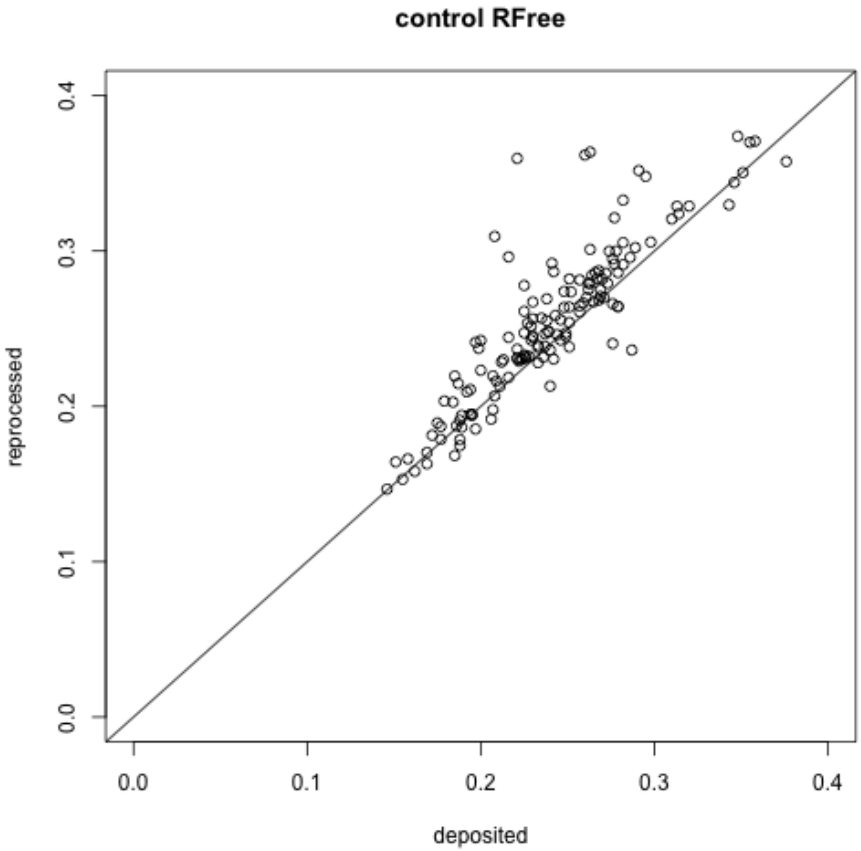
What to do with new metadata?

- “Correct” images
  - Complicates data integrity
  - Headers not designed for multiple sources of information
- REST API
  - Complicates analysis pipelines


20160516
Intensities 83%
Failed 17%


20160516, API assisted
Intensities 91%
Failed 9%


reprocessable
PG match
SG match
1.0
0.8
0.6
0.4
0.2
0.0
NC2016
20170516
20170516 API assisted


control RFree
reprocessed
deposited
0.0
0.1
0.2
0.3
0.4

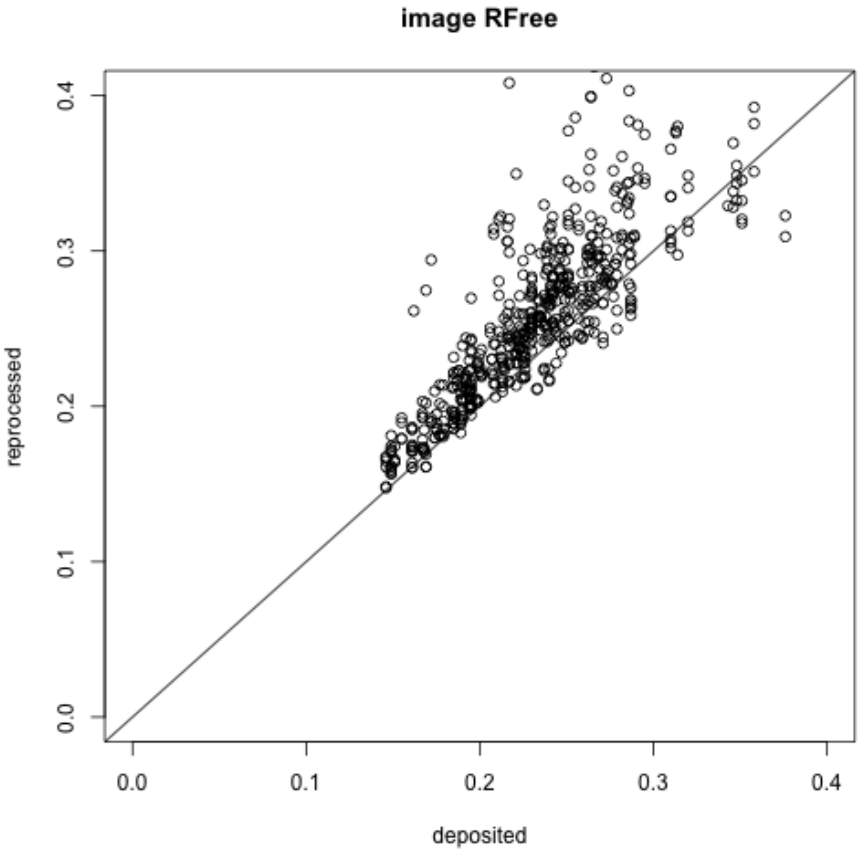


image RFree
reprocessed
deposited
0.0
0.1
0.2
0.3
0.4

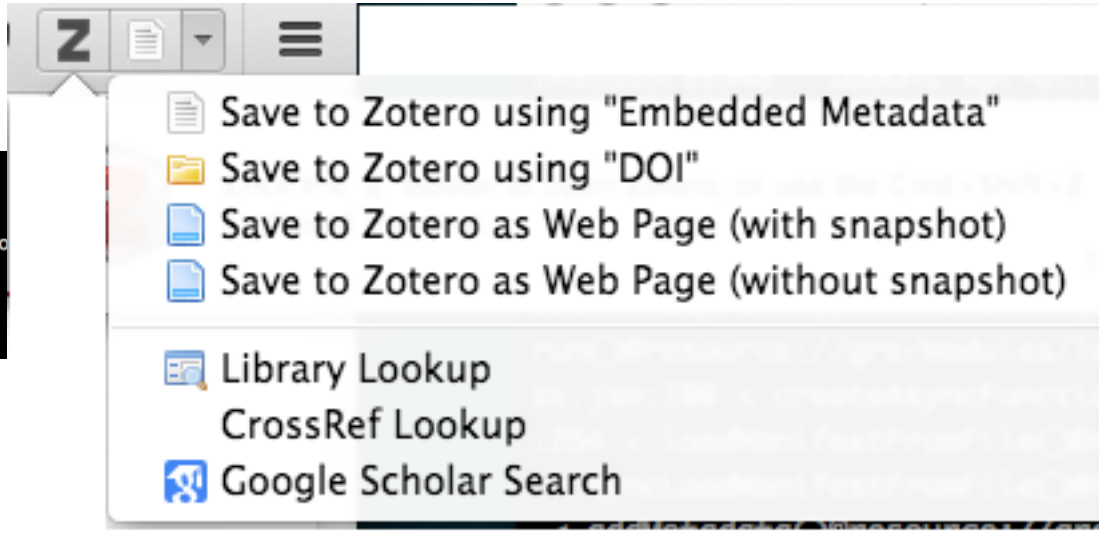
Connecting to Scientific Ecosystem

Cite this Dataset

Nam, Y; Sliz, P. 2015. "X-Ray Diffraction data for: Lin28A/let-7g microRNA complex. PDB Code 3TS2", SBGrid Data Bank, V1, http://dx.doi.org/10.15785/SBGRID/1.

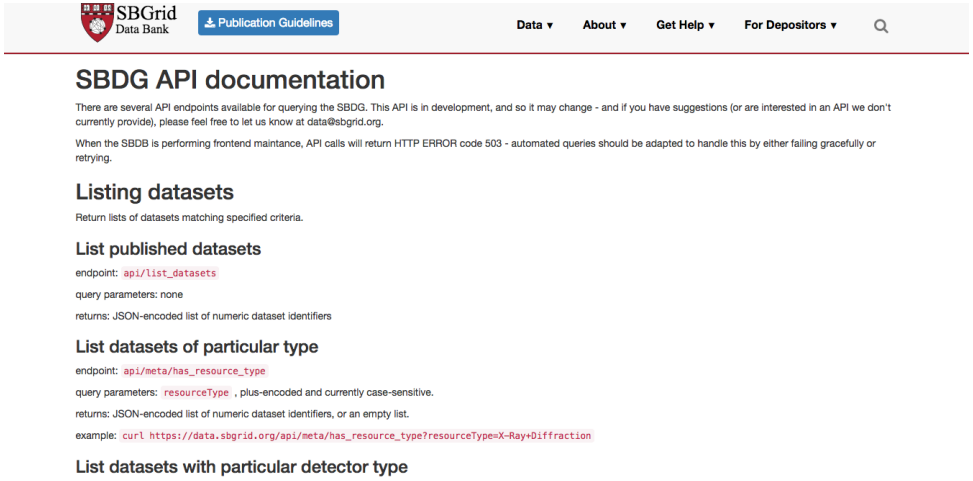
Download Citation

```
@data{ sbdb1,
 author = { Nam, Y and  Sliz, P   },
 publisher = { SBGrid Data Bank },
 title = { X-Ray Diffraction data for: Lin28A/let-7g microRNA complex. PDB Co
 year = { 2015 },
 doi = { 10.15785/SBGRID/1 },
 url = { http://dx.doi.org/10.15785/SBGRID/1 }
}
```



Connecting to Scientific Ecosystem

- Integrations
  - RCSB integration
  - DataMed
- API



https://data.sbgrid.org/static/html/apidoc/index.html

What's next?

- Dataverse integration
  - Improved UI
  - ORCID login
- Improved analysis pipelines
  - Additional tools
  - API generalization
- User suggestions

Acknowledgements

Stephanie Socias
Jason Key
Bill McKinney
Mick Timony
Justin O'Connor
Michelle Ottaviano
Piotr Sliz

Phil Durbin
Gustavo Durand
Leonid Andreev
Mercè Crosas

Developers / Projects:
- *xia2*
- CCP4
- XDS
- MOSFLM
- PHENIX
- LABELIT

Crystallographers

Alejandro Buschiazzo
Ming Lei
Filipe Maia

Protein Data Bank

Orchestra
HIGH PERFORMANCE COMPUTING CLUSTER

THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST