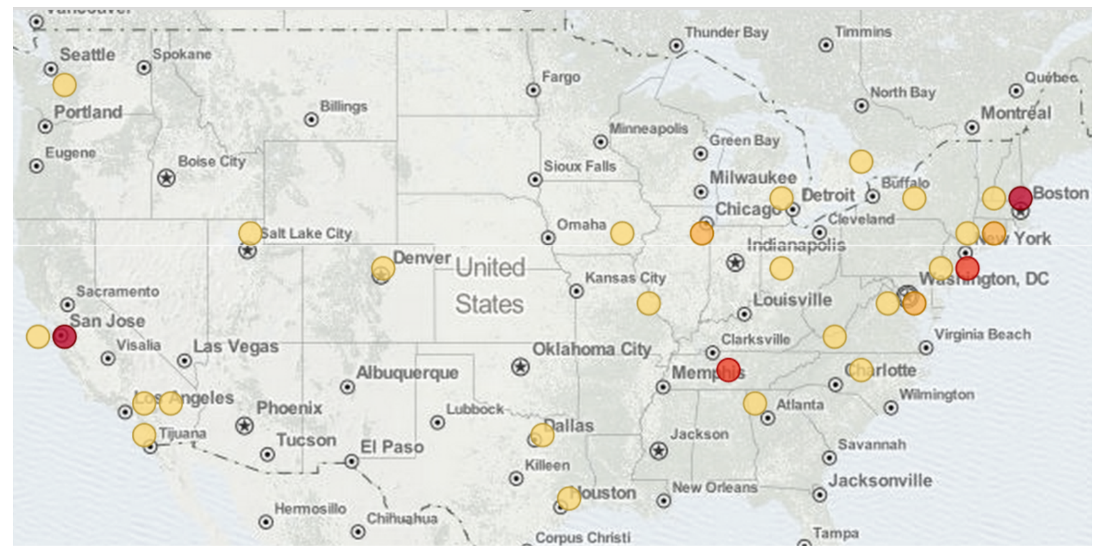


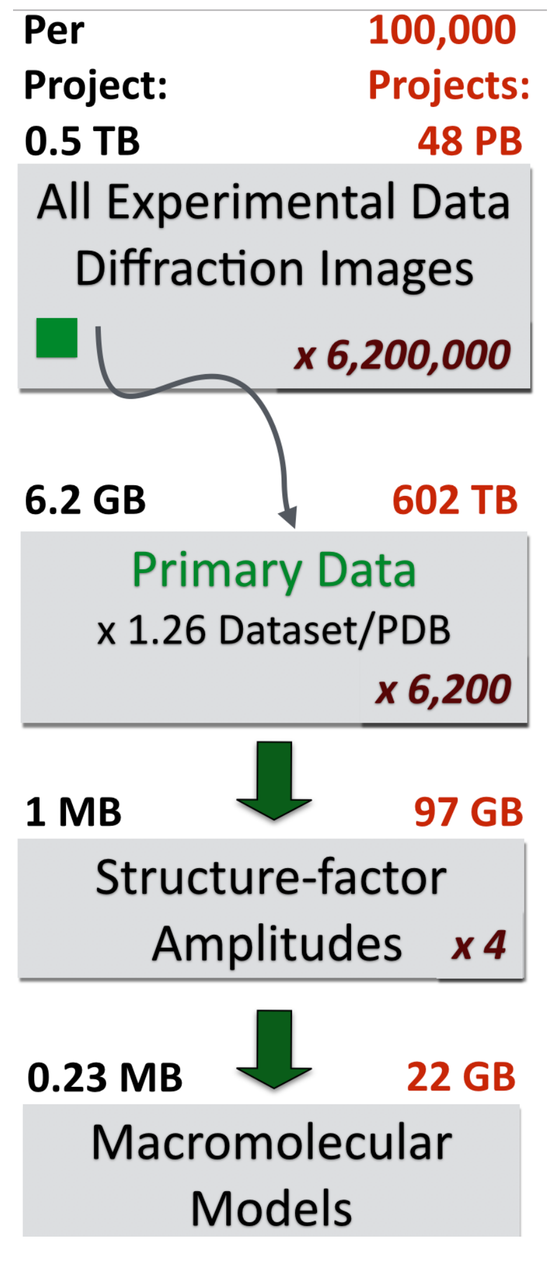
Data Publication and Dissemination with the Structural Biology Data Grid

Stephanie Socias, Peter Meyer, Emily Tjon, David Oh, Jiawei Wu, Mercè Crosas[#], Piotr Sliz

SBGrid Consortium and [#]Dataverse, Harvard University

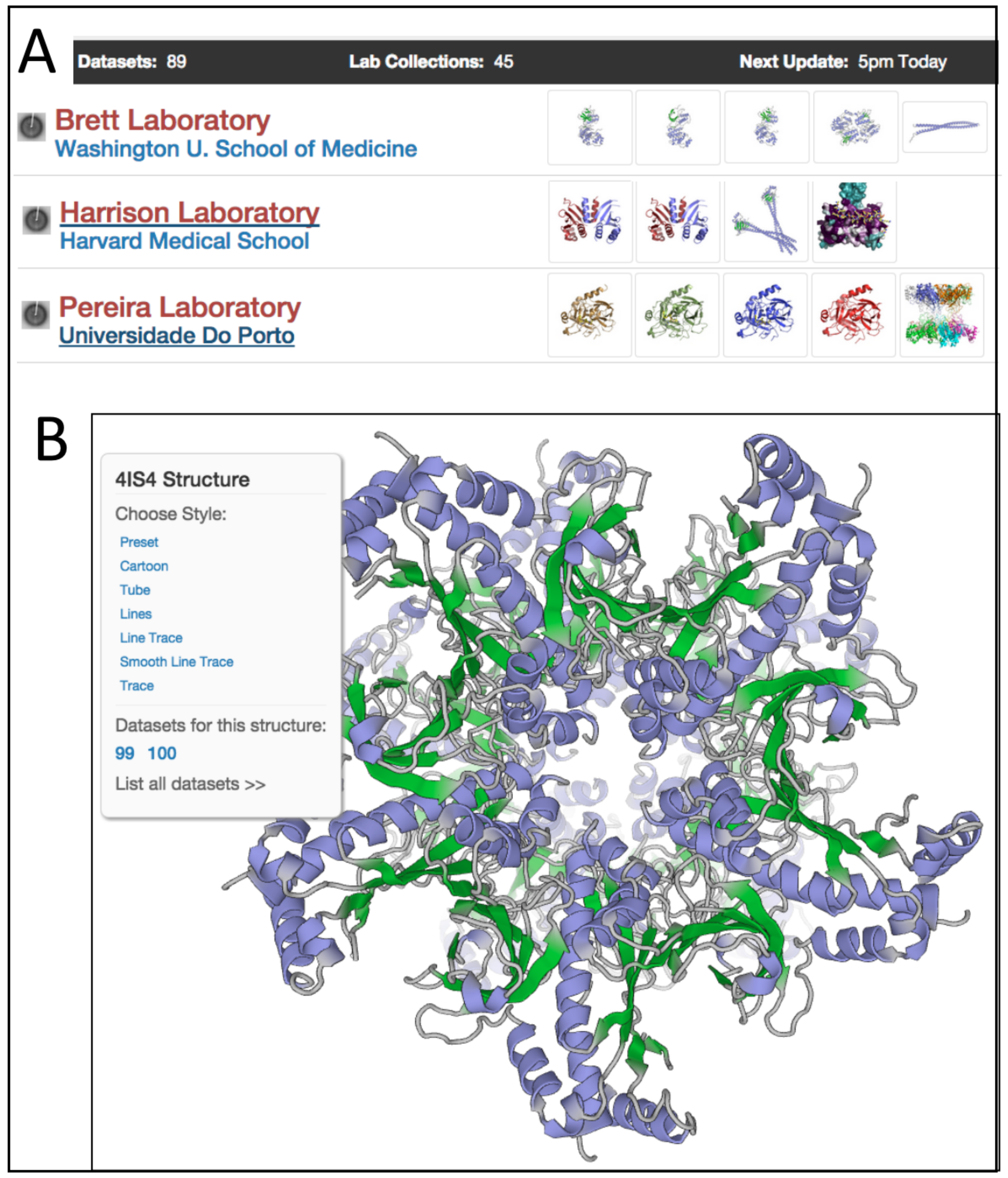
Abstract:

Access to experimental X-ray diffraction image data is fundamental for validation and reproduction of macromolecular models and indispensable for development of structural biology processing methods. In response to evolving needs of the structural biology community, we established a diffraction data publication and dissemination system, Structural Biology Data Grid (SBDG, url: data.sbgrid.org), to preserve primary experimental datasets that support journal publications. Datasets archived with the SBDG are freely available to the research community under a public domain dedication license and the metadata for all datasets is published under the DataCite schema. Datasets are accessible to researchers through the Data Access Alliance infrastructure, which facilitates global and institutional data access. Our analysis of a pilot collection of crystallographic datasets demonstrates that the information archived by SBDG is sufficient to reprocess data to statistics that meet or exceed the quality of the original published structures. It is anticipated that access to the experimental datasets will enable paradigm shift in the community from the static archive towards a much more dynamic body of continuously improving refined models. Following the success of this pilot study, the SBDG has extended its services to the entire community and will be used to develop support for other types of biomedical datasets, such as MicroED, Molecular Dynamics trajectories and Lattice Light-Sheet Microscopy.



Estimation of Storage requirements for different stages of the structural biology pipeline, based on the SBDG pilot collection.

Website: The SBDG's collection of datasets can be accessed from the data.sbgrid.org website. On the home page, deposited datasets are organized into laboratory and institutional collections. Hyperlinked collection pages provide a list of selected datasets along with the dataset's corresponding data Digital Object Identifier (DOI), a link to the journal publication, the PDB ID, a link to the PDB entry, and a link to the depositors' laboratory website. The website molecular viewer, PV offers visitors an option to view structures in a manipulatable cartoon representation. The website is being migrated to the Dataverse open-source software (http://dataverse.org), which provides a rich set of features and best practices for an open data repository.



Datasets: 89 Lab Collections: 45 Next Update: 5pm Today

Brett Laboratory
Washington U. School of Medicine

Harrison Laboratory
Harvard Medical School

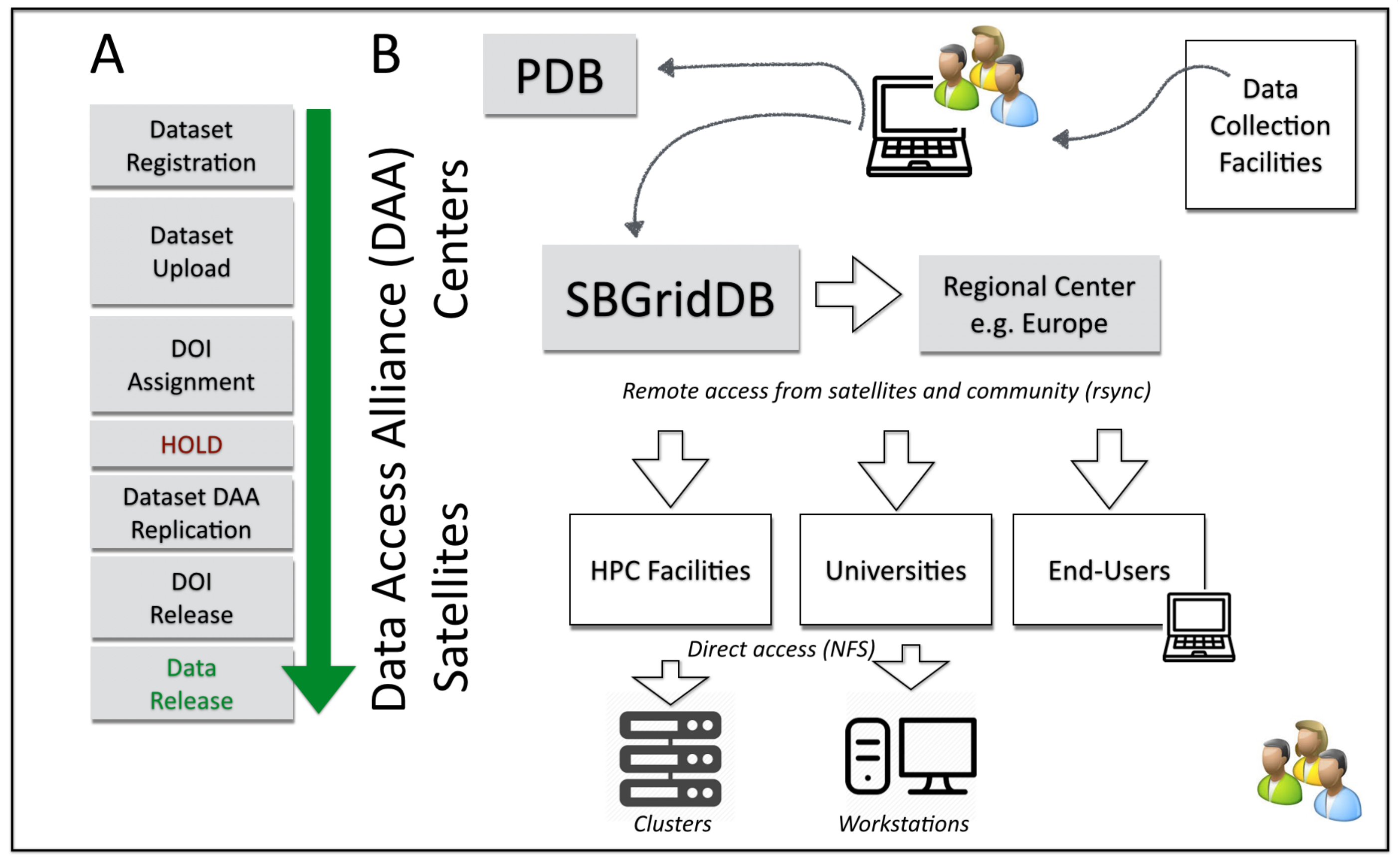
Pereira Laboratory
Universidade Do Porto

4IS4 Structure

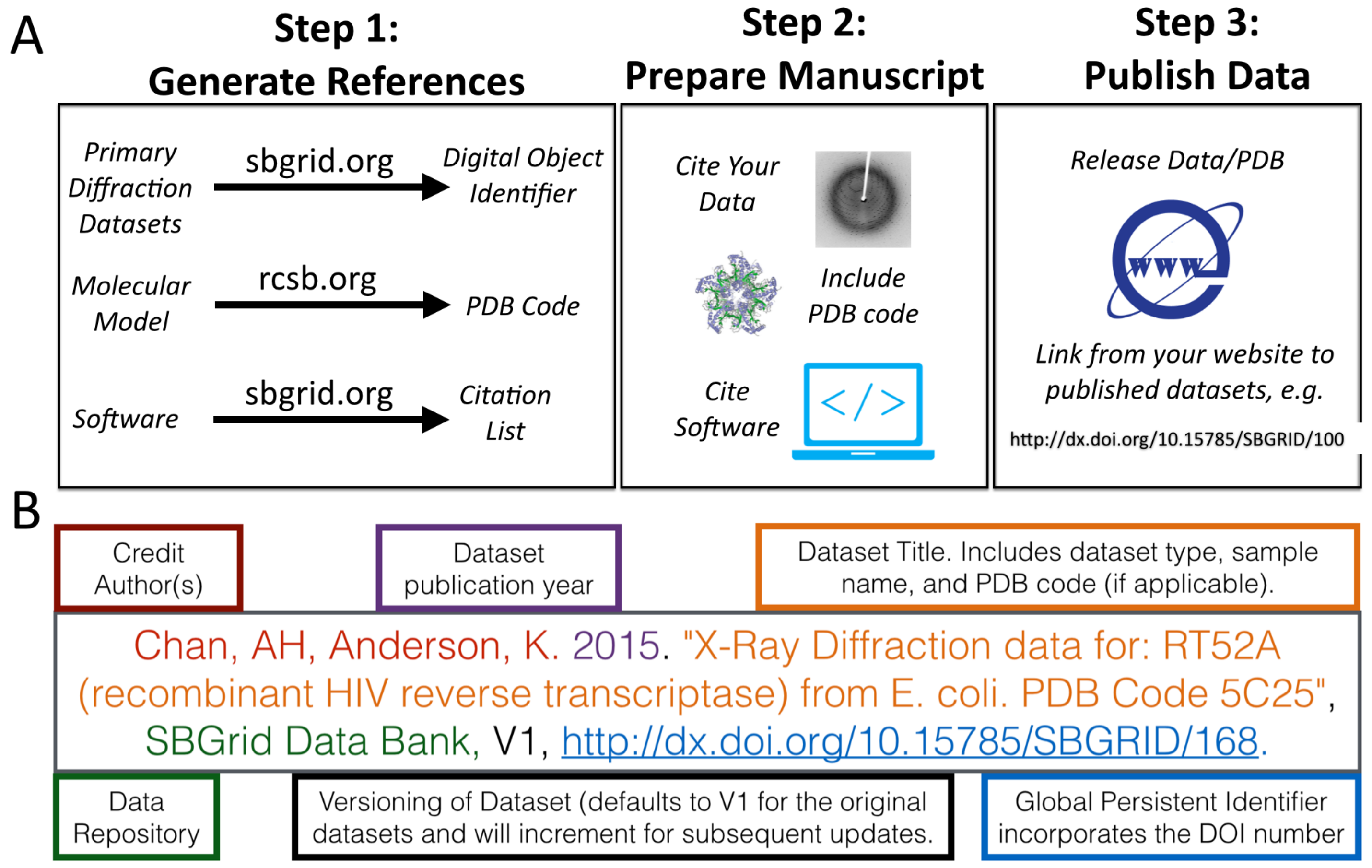
Choose Style:
Preset
Cartoon
Tube
Lines
Line Trace
Smooth Line Trace
Trace

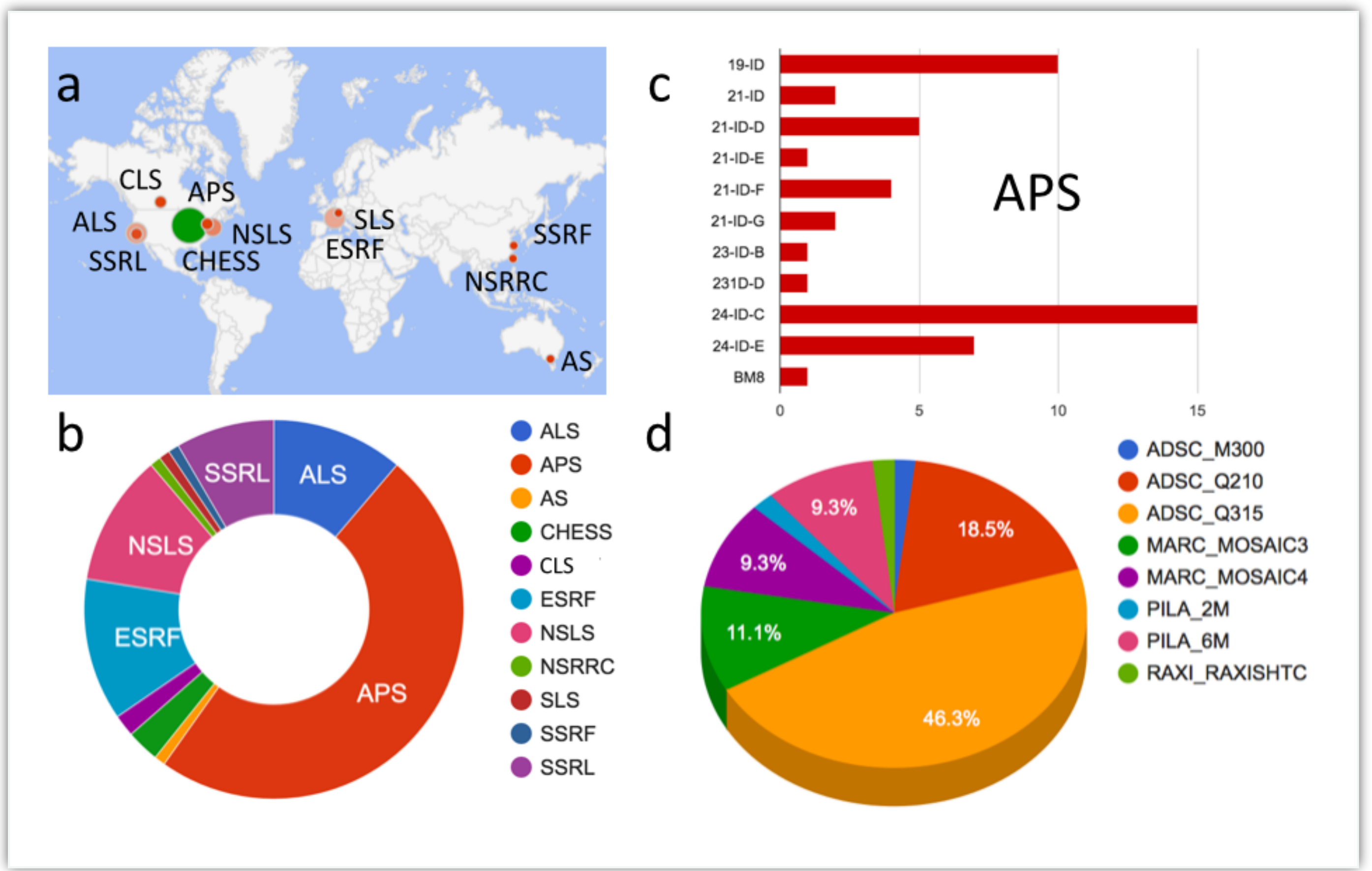
Datasets for this structure:
99 100
List all datasets >>

Data Grid: Physical access to SBDG datasets is facilitated through a data grid infrastructure that is supported by members of the Data Access Alliance (DAA, Fig. 5). The DAA is an open organization of research-data-storage providers, being developed in collaboration with the Globus Project as part of a National Data Service (NDS) pilot.

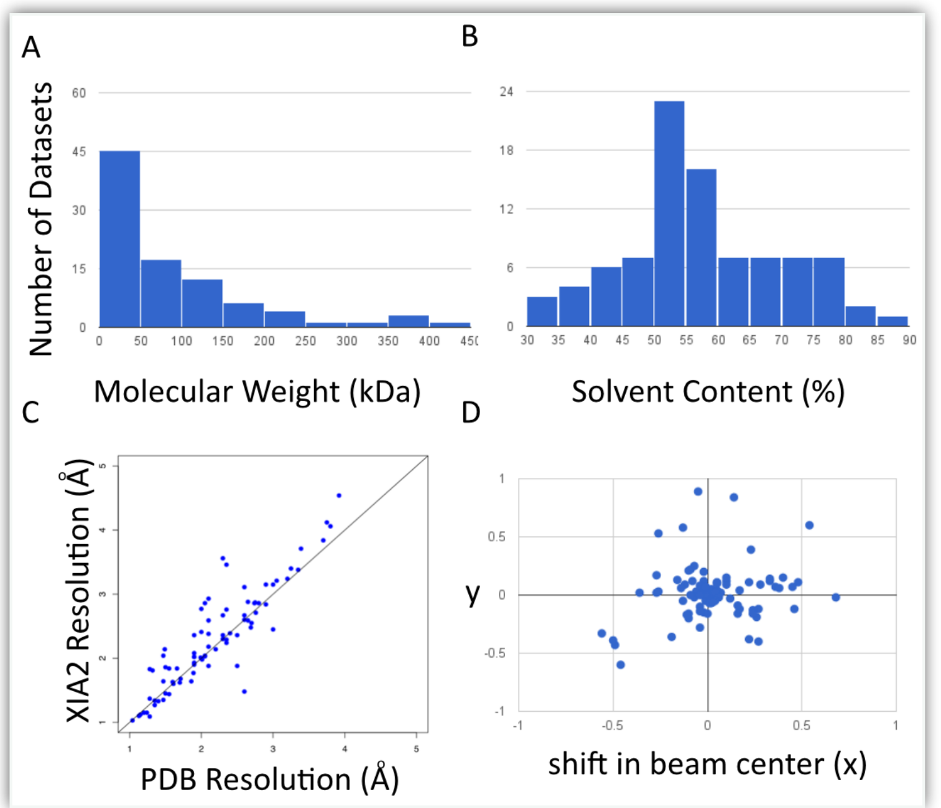


Data Publication: Research data are legitimate and citable products of research (Bourne et al., 2012) and, therefore, the SBDG recommends that depositors and data users cite all data deposited with the SBDG in the standard reference section of their manuscripts following well established community standards (Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone (ed.) San Diego CA: FORCE11; 2014).


A
Step 1: Generate References
Primary Diffraction Datasets
sbgrid.org
Digital Object Identifier
Molecular Model
rcsb.org
PDB Code
Software
sbgrid.org
Citation List
Step 2: Prepare Manuscript
Cite Your Data
Include PDB code
Cite Software
Step 3: Publish Data
Release Data/PDB
WWW
Link from your website to published datasets, e.g.
http://dx.doi.org/10.15785/SBGRID/100
B
Credit Author(s)
Dataset publication year
Dataset Title. Includes dataset type, sample name, and PDB code (if applicable).
Chan, AH, Anderson, K. 2015. "X-Ray Diffraction data for: RT52A (recombinant HIV reverse transcriptase) from E. coli. PDB Code 5C25", SBGrid Data Bank, V1, http://dx.doi.org/10.15785/SBGRID/168.
Data Repository
Versioning of Dataset (defaults to V1 for the original datasets and will increment for subsequent updates.
Global Persistent Identifier incorporates the DOI number



Data collection statistics for the pilot subset of 110 datasets. (a,b) Datasets were collected from synchrotrons on four continents (in addition to laboratory sources, which are not broken down geographically) and originate from eleven synchrotron facilities. Datasets cover a range of detector types, including Area Detector Systems Corporation M300, Q210 and Q315, Rayonix MarMosaic, Dectris Pilatus 2M and 6M, R-AXIS HTC, and MAR345.



Data Quality: For a proof of concept, released datasets in the SBGridDB were reprocessed with XIA2 in a fully automated manner. 90 of the 110 released datasets with a corresponding PDB ID were successfully reprocessed. 86 of those 90 datasets represented high-resolution, native data and for 51 of those XIA2 automatic data processing arrived with CC1/2 resolution within 0.1 Å of the published structure

Teaching Collection. 12 X-ray diffraction datasets from the SBDG pilot collection were identified as particularly suitable for software testing and teaching activities.

10.15785/SBGRID/5 Boggon Laboratory	Datasets from 5 crystals of SNX17 FERM domain in complex with a peptide corresponding to KRIT1's NPxY2 motif. Separate integration of the datasets and scaling together allows a complete 3.0 Å dataset for molecular replacement solution (original paper used 4GXB as a search model) and structure refinement.
10.15785/SBGRID/117 Baxter Laboratory	3.70 Å dataset collected on a crystal of thioester-containing protein 1 *S1 allele (TEP1*S1). Initial data processing suggested P43212, but one of the two molecules (~1300 aa. each) in the ASU overlapped with it's symmetry-mate. Comparison of alternative scenarios in refinement identified the true space group as *P*4_3 with twinning and rotational pseudosymmetry. Refinement was completed with TLS, NCS (local) and external restraints derived by *ProSMART* using TEP1*R1 (PDB 4D94) as reference.
10.15785/SBGRID/62 Modis Laboratory	4.5 Å dataset of a uranyl acetate derivative used for a challenging structure determination by SAD. Certain images had streaky features and were excluded from data reprocessing. The height and definition of peaks in anomalous difference Patterson maps was improved by omitting certain images near the end of the data collection run.
10.15785/SBGRID/111 Ferré-D'Amaré Laborator	2.5 Å dataset collected at ALS BL 5.0.2 using 6.0 keV X-rays from a crystal of 'Spinach' a fluorescent RNA analog of GFP. Although anomalous signal was very weak, a heavy atom substructure comprised of one barium and six potassium ions resulted in good quality SAD electron density maps.
10.15785/SBGRID/3 Sliz Laboratory	2.9 Å Zn SAD dataset was sufficient to determine a crystal structure of Lin28/let-7d protein-microRNA complex. X-ray beam size was adjusted to maximize flux and minimize radiation damage. One swapped-dimer is located in each asymmetric unit. Two native zinc atoms are located in each tandem CCHC zinc knuckles domain.
10.15785/SBGRID/123 Heldwein Laboratory	This 3.29-Å selenomethionine SAD data set, collected at 0.9789 angstrom wavelength at BNL X25 beamline, was sufficient to determine the phases and to trace the structure of HSV-2 gH/gL complex (3M1C, Chowdary et al, 2010). There are 9 Se sites in the ASU. During integration in HKL2000, chi2 appeared very large for some sectors of the data set. These correlated with crystal orientation and likely resulted from a large difference in cell edges (88 Å vs 333 Å).
10.15785.SBGRID/179 Schwartz Laboratory	Contaminating *E.coli* protein 4FCC_A, acting as a crystallization chaperone, was found readily by MR. Using these MR phases 7 $(Ta_6Br_{12})^{2+}$-positions could be found in the 8.8 Å derivative dataset 180. The combined MR-SAD phases were sufficient to position two copies of Nup37 (4FHL) and two copies of Nup120 in the asymmetric unit.
10.15785/SBGRID/78 Rudenko Laboratory	2.65 Å data set collected at APS using multiple settings on a crystal of the neurexin 1alpha ectodomain. The structure has 2 molecules/ASU with a total of 14 ordered domains and ~2000 residues. Molecular replacement successfully placed 8 LNS domains (using a single LNS domain as a search model, i.e. ~9% of the scattering mass) generating phases which could be used to reveal 37 out of 44 Se atoms/ASU in a 3.25 Å SeMet SAD data set.
10.15785/SBGRID/9 Tao Laboratory	A 3.25-Å resolution dataset was collected at APS LS-CAT. The structure was determined by molecular replacement using a 9-Å resolution cryo-EM reconstruction as a phasing model. Solvent flattening and 15-fold noncrystallographic symmetry averaging were applied during phase extension (Crystal structure of a nematode-infecting virus. Guo, Hryc, Jakana et al, PNAS, 2014).
10.15785/SBGRID/83 Drennan Laboratory	Diffraction data from different regions of a crystal of Isobutyryl-coenzyme A mutase fused, a 250 kDa dimeric enzyme. This crystal had a large unit cell (a,b = 319 Å, c = 344 Å) and the data were anisotropic. Separate integration of the 6 wedges with individually adjusted resolution limits and scaling together yields a complete 3.35 Å dataset that can be used for molecular replacement.
10.15785/SBGRID/125 Kruse Laboratory	Diffraction data for crystals lipidic cubic phase of human M_2 muscarinic acetylcholine receptor bound to the agonist iperoxo, the allosteric modulator LY2119620, and the conformationally-selective nanobody Nb9-8.
DOI: 10.15785/SBGRID/68	1.2 Å dataset collected at SSRL provides a high-resolution standard dataset of the enzyme Cyclophilin to examine the influence of data collection temperature to compare to XFEL data, and to measure X-ray diffuse scattering.

Summary: Access to this growing collection of X-ray diffraction datasets will support the proposed paradigm shift in the community (Terwilliger and Bricogne, 2014) from the static archive towards a much more dynamic body of continuously improving refined models.