

## MIT Open Access Articles

*Data publication with the structural biology data grid supports live analysis*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Meyer, Peter A.; Socias, Stephanie; Key, Jason; Ransey, Elizabeth; Tjon, Emily C.; Buschiazzi, Alejandro; Lei, Ming, et al. "Data Publication with the Structural Biology Data Grid Supports Live Analysis." Nature Communications (March 2016): 10882.

**Published Version:** <http://dx.doi.org/10.1038/ncomms10882>

**Publisher:** Nature Publishing Group

**Permanent Link:** <http://hdl.handle.net/1721.1/108692>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of use:** <http://creativecommons.org/licenses/by/4.0/>



ARTICLE

Received 16 Oct 2015 | Accepted 28 Jan 2016 | Published 7 Mar 2016

DOI: 10.1038/ncomms10882

OPEN

# Data publication with the structural biology data grid supports live analysis

Peter A. Meyer *et al.*<sup>#</sup>

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. Here, we established a diffraction data publication and dissemination system, Structural Biology Data Grid (SBDG; [data.sbgrid.org](http://data.sbgrid.org)), to preserve primary experimental data sets that support scientific publications. Data sets are accessible to researchers through a community driven data grid, which facilitates global data access. Our analysis of a pilot collection of crystallographic data sets 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. SBDG has extended its services to the entire community and is used to develop support for other types of biomedical data sets. It is anticipated that access to the experimental data sets will enhance the paradigm shift in the community towards a much more dynamic body of continuously improving data analysis.

---

Correspondence and requests for materials should be addressed to P.S. (email: [sliz@hkl.hms.harvard.edu](mailto:sliz@hkl.hms.harvard.edu)).

<sup>#</sup>A full list of authors and their affiliations appears at the end of the paper.

As one of the most powerful tools in structural biology, X-ray crystallography allows determination of the structure (atomic coordinates) of proteins, nucleic acids, small molecule compounds and macromolecular complexes to atomic-level resolution. Crystallographic data continue to be a primary source of mechanistic understanding of macromolecules, the implications of which extend from basic research to translational studies and the rational design of therapeutics. Reflecting the significance of the technique, the number of published macromolecular crystal structures has rapidly grown to >100,000 and numerous investigators within structural biology have been awarded the Nobel Prize, including Drs. Kendrew, Perutz, Watson, Crick, Wilkins, Hodgkin, Klug, Deisenhofer, Michel, Huber, Walker, MacKinnon, Kornberg, Ramakrishnan, Steitz, Yonath and Kobilka.

To support the needs of a growing structural biology community, a global network of synchrotron beamlines<sup>1</sup> has been established and made available to researchers. These facilities remain the predominant source for crystallographic data collection. While the data collection process has become increasingly streamlined, deployment of a data management infrastructure to archive original diffraction images has been slow and uncertain<sup>2</sup>. With the exception of a modest number of data storage systems dedicated to the support of individual synchrotron beamlines<sup>3</sup>, or specific structural genomics projects<sup>4</sup>, storage of diffraction image data sets is typically the responsibility of primary investigators. Access to these original experimental data sets is therefore dependent on the policies of individual laboratories, which vary in storage organization, institutional resources, and researcher turnover. There is no universal archiving system to store X-ray diffraction data sets, and raw data sets are rarely made publicly available. In the cases where data sets are available, their distribution format can vary significantly. A typical data set of 360 images collected on modern detectors is 5 GB, and structure determination can involve one to tens of data sets, making the logistics of storing diffraction data for many protein structures a daunting task.

The benefits of easy and public access to experimental data are numerous<sup>5</sup>. Access to primary data would support community efforts to continuously improve existing models and identify new features through complete reprocessing of experimental data<sup>6–8</sup> with modern software tools and improved criteria<sup>9</sup>. Further, original data may provide a basis for validating questionable existing structures while mistakes in structure determination may be identified earlier<sup>10–12</sup>. Additionally, access to a diverse volume of raw data can be used to develop improved software to address limitations of existing programs. Finally, access to a collection of varied experimental data will undoubtedly benefit the training and education of practitioners. The Worldwide Protein Data Bank<sup>13,14</sup> (wwPDB) has illustrated how these achievements can be realized with the collection of reduced experimental data, in the form of structure factor amplitudes. Complementing this resource by preserving raw experimental data and making it available to a broad community promises a profound scientific impact in structural biology and other biomedical disciplines that face the challenges of preserving large data sets.

While the primary role of the SBGrid Consortium ([www.sbgrid.org](http://www.sbgrid.org)) has been to curate and support a collection of data processing software applications and to organize community-wide computing support<sup>15</sup>, SBGrid has also been active in the management of raw, experimental data sets. In 2012, SBGrid prototyped a system based on Globus technology<sup>16–19</sup> to move diffraction data between Harvard, The Advanced Photon Source, and the Stanford Synchrotron Radiation Light source<sup>19</sup>.

To support the outstanding needs of the global structural community, we have established a publication system for

experimental diffraction data sets that supports published structural coordinates: the Structural Biology Data Grid (SBDG). The SBDG project was initiated with a collection of X-ray diffraction image data sets as well as a few additional data set types contributed by many SBGrid Consortium laboratories. The collection supports a diverse subset of over 68 peer-reviewed publications and represents a sampling of numerous structure determination approaches. To evaluate the utility of such a data grid, we reprocessed all published diffraction data sets in this initial collection with modern software and compared the derived statistics against those reported in the original publications. We also demonstrate that by integrating the storage resources of multiple research groups and institutions, the data grid is poised to deliver a novel community driven data preservation system to support various types of structural biology and biomedical data sets.

## Results

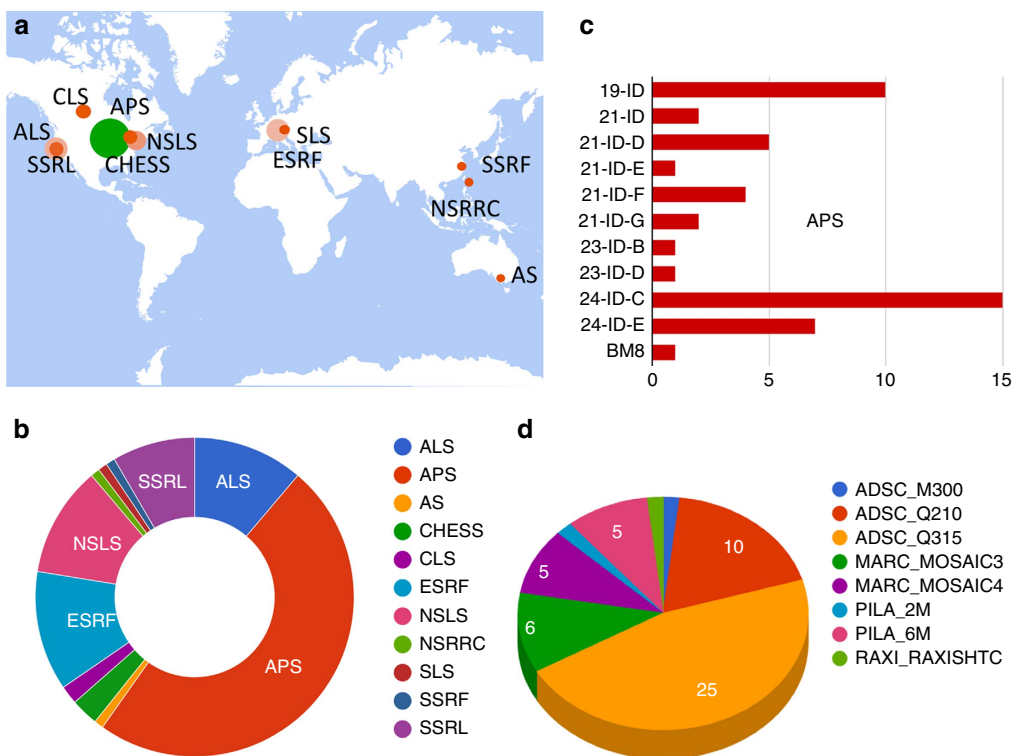
**Structural biology data grid.** The SBDG is a centralized data publication service—a repository for discovering, downloading and depositing large structural biology data sets. We developed the SBDG to support the need of the SBGrid community to archive and disseminate X-ray diffraction image data sets, that is, images recorded on X-ray detectors, which support published structures. More than 90% of SBGrid laboratories use X-ray crystallography in their research, and SBGrid investigators have contributed over 11,000 X-ray structures to the PDB. The SBDG complements the PDB, which archives derived data—merged and post-refined data from diffraction images and the resulting refined coordinates of macromolecular structural models. The data grid has been developed in collaboration with the Data Science team at Harvard's Institute for Quantitative Social Science, and it conforms to progressive data science standards (Table 1). The SBDG limits its collection to data sets that support journal publications, referred to as 'primary data'. For X-ray diffraction data, this primary data consists of experimental diffraction images supporting a derived structural model and journal publication. Release of this primary data by the SBDG coincides with publication of the resulting manuscript and for the structural biology data sets of related PDB files. As of 1 September 2015, the SBDG stores a diverse collection of 117 data sets, including 111 X-ray diffraction data sets and a handful of other data types including computational decoys and data sets from MicroED, lattice light-sheet microscopy and molecular dynamics (Supplementary Table 1). These published data sets, contributed by 50 laboratories with diffraction data sets collected at 11 synchrotron facilities (Fig. 1) and several home sources, originated 94 structures and 68 journal publications. The X-ray diffraction data sets range in size from 126 MB (ref. 20) to 20 GB (ref. 21) with a mean of 4.9 GB and a total of 573 GB of storage. Extrapolating from this initial collection, which is quite diverse and registers at just over 0.5 TB, our current 100 TB file system could immediately support roughly twenty thousand X-ray diffraction data sets (Fig. 2).

The SBDG's collection of data sets can be accessed from the [data.sbgrid.org](http://data.sbgrid.org) website. On the home page, deposited data sets are organized into laboratory and institutional collections (Fig. 3a). Hyperlinked collection pages provide a list of selected data sets along with the data set'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<sup>22</sup>, offers visitors an option to view structures in a manipulatable cartoon representation (Fig. 3b). With multiple high-quality viewing options and flexible search functionality, users of the SBDG website can easily identify a small subset of relevant data sets.

**Table 1 | Data science standards.**

Disclosure	Software tools developed under this program will be incorporated into open source software and released to the community. Manuscripts and white papers describing various phases of the project will be released on a regular basis.
Adoption	All biomedical image data will be converted to the master formats, such as OME-TIFF or HDF5. Community tools to create, analyse, and manipulate diffraction images will be extended to include support for these formats. All biomedical data are assigned Digital Object Identifiers through the CDL EZID system, and follow modified DataCite and Dataverse metadata schemas. Associated metadata are registered with the International DOI Foundation, making it virtually permanent and independent of SBDG and Harvard computing infrastructure. All data sets published through the SBDG will be citable using Force 11 recommendations.
Transparency	Files within individual data sets will be deposited in their original format (no archives or encryption allowed). Self-documentation: The majority of diffraction data sets are self-documented and include the basic information required for reprocessing in the images themselves. Additional information will be collected during deposition and will include data set representation (the ability to use the data to be processed), reference (relation to PDB files, publications, and other data sets), context (for example, a native data set or a derivative used for phasing), fixity (checksums), and provenance (typically the data collection facility and the project member who deposits the original data set). With conversion to master formats, all secondary information will be appended to the image metadata along with all original headers.
External dependencies	The ability to reprocess some older data sets and verify master format conversions could depend on access to a specific version of data processing software. As data sets enter our repository, they will be reprocessed with our Data Reprocessing Pipeline (one of several we will develop as part of our Data Mining Pipelines). Data Reprocessing Pipelines will be archived within our system, issued DOIs, and interlinked with the data sets. It is worth noting that, since 2002, SBDG has been archiving structural biology applications and, therefore, has access to previous software versions that might be required to reprocess older data sets.
Licensing	Biomedical data sets will be deposited under the Creative Commons Zero licence, supporting future development of data validation services and database replications and migrations.
Technical protection mechanism	The security of the deposited data will be maintained by the DAA. The DAA will join with the Library of Congress sponsored NDSA and the data architect working on the project will ensure that NDSA recommendations are being followed.

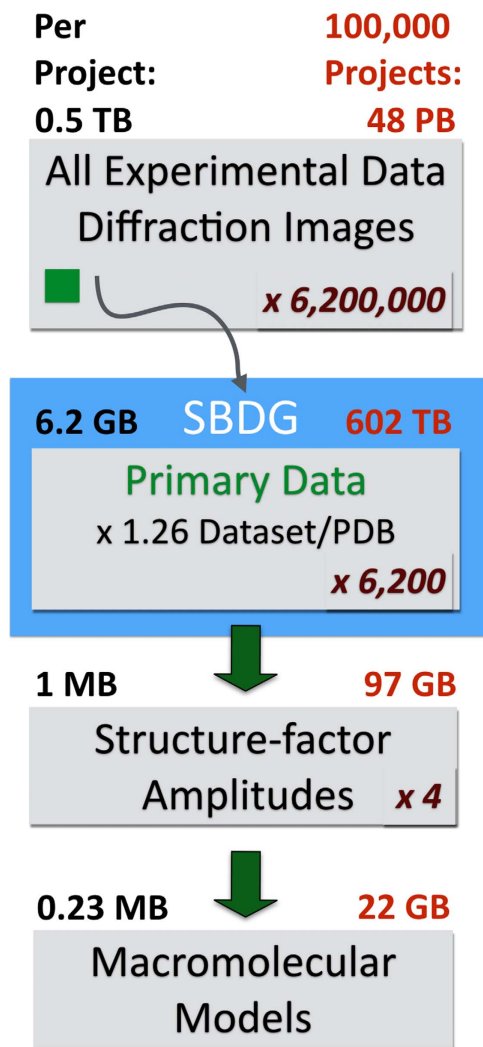
NDSA, National Digital Stewardship Alliance; SBDG, Structural Biology Data Grid.



**Figure 1 | Data collection statistics for the pilot subset of 112 data sets.** (a,b) Data sets were collected from synchrotrons on four continents (in addition to laboratory sources, which are not broken down geographically) and originate from eleven synchrotron facilities: Advanced Light Source, Advanced Photon Source, Australian Synchrotron, Cornell High Energy Synchrotron Source, Canadian Light Source, European Synchrotron Radiation Facility, National Synchrotron Light Source, National Synchrotron Radiation Research Center, Swiss Light Source, Shanghai Synchrotron Radiation Facility, and Stanford Synchrotron Radiation Lightsource. World map image courtesy of the U.S. Geological Survey. (c) Breakdown of data sets collected at the Advanced Photon Source beamlines. (d) Data sets 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.

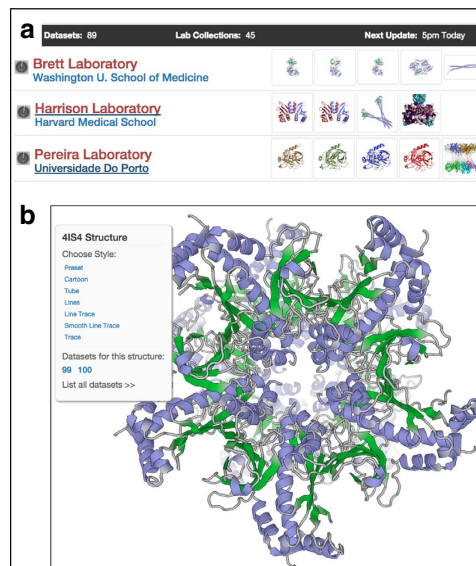
Persistent data set pages are an important element for any research data repository because they typically provide a landing URL, which resolves from a given DOI<sup>23</sup>. The SBDG does not

advertise unique codes, but instead distinguishes data sets by fully qualified DOIs. From each SBDG collection page or viewer page a user can access those unique Data set Pages (Fig. 4), which offer



**Figure 2 | Estimation of storage requirements for different stages of the structural biology pipeline, based on the SBDG pilot collection.** For structure factor amplitudes and PDB models file sizes were obtained from a subset of 96 PDB depositions derived from the pilot data sets. On average, SBDG stores 1.26 data sets per PDB file. Numbers in red indicated the estimated storage requirements to accommodate data sets for 100,000 structures. We estimate that for each primary data set, additional 100 data sets are collected at national facilities. Primary data refers to original experimental diffraction images supporting the derived structural model, as distinguished from all experimental data (screening images, inferior quality data sets, and so on). For crystallographic experiments, reduced data refers to the integrated intensities (or amplitudes, which do not materially affect storage requirements).

additional information for each data set including download instructions and the fully formatted data set citation for inclusion in manuscripts, following best practices set by the Joint Declaration of Data Citation Principles<sup>24</sup>. A Data set Page can also be located by searching the SBDG for a PDB code, although often several related data sets are used to determine a single set of macromolecular coordinates. As the Data Grid is developed, the Data set Pages will include additional functionality, with more information on how to reprocess data sets, extended data statistics, and discussion forums allowing users to annotate data sets after publication. Taken together, the uniquely defined Data set Pages provide a comprehensive and persistent location for individual data sets.



**Figure 3 | Organized display of data collections at SBDG.** (a) Graphical view of Laboratory and Institutional Collections within the SBDG; (b) PV structure viewer, displaying a published model with links to its two primary deposited data sets.

**Data set access.** All data sets in the SBDG are readily and freely accessible to the community. Access rights were formalized with adoption of the creative commons zero licence (CC0), which supports dedication of research results to the public domain and is used by many open-data projects. This licence allows use and redistribution of data for both commercial and non-commercial purposes without requiring additional agreements. The CC0 licence does not affect patents or trademark rights of contributors, and is similar to the licensing terms that are used for macromolecular models released by the wwPDB.

Although data sets can be downloaded individually, their size can make this cumbersome. Physical access to SBDG data sets is facilitated through a data grid infrastructure that is supported by members of the data access alliance (DAA; Fig. 5a). The DAA is a voluntary and open organization of research-data-storage providers and is being developed in collaboration with the Globus Project. The DAA has two aims: (1) to minimize the chance of data loss by replicating SBDG data sets, and (2) to facilitate global data access through its members. Although it is expected that DAA membership and architecture will evolve rapidly, in its current state the DAA framework already provides a global solution for data dissemination. DAA centres in Europe, Asia, North America and South America replicate the entire SBDG collection and provide local access to members of regional communities. There are four DAA centres: Harvard Medical School in the USA, Uppsala University in Sweden, Shanghai Institutes for Biological Sciences in China, and Institut Pasteur de Montevideo in Uruguay. As a secondary service, DAA centres can provide local, direct access to data sets for their institutional research groups. For example, Harvard Medical School hosts the entire collection and provides direct access to all data from its computing center. The DAA infrastructure is further extended by the DAA satellites, which replicate fractions of SBDG data sets in their local storage for direct access by members of individual institutions. This mode of participation provides an attractive option for research institutions to develop local archives of all primary data generated by the local community. For example, the NE-CAT (Northeastern Collaborative Access Team; sector 24-ID) synchrotron beamline at the Advanced Photon Source, in

**SBGrid DATA BANK** For Depositors Data About More

**X-Ray Diffraction data from *Medicago truncatula* glutamine synthetase, source of 4IS4 structure**

Data DOI: 10.15785/SBGRID/100 | ID: 100  
 Publication DOI: 10.1107/S1399004713034718  
 4IS4 Coordinates: [Viewer](#), [PDB](#), [MMDB](#)  
[Pereira Laboratory](#), Universidade Do Porto  
 Release Date: May 19, 2015

**Download Instructions**

- 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/100.'
```
- After the transfer is completed, please issue the following command to verify data integrity:  

```
'cd 100 ; shasum -c files.sha'
```

**Biological Sample:**  
*Medicago truncatula* glutamine synthetase  
**Dataset Type:**  
 X-Ray Diffraction  
**Subject Composition:**  
 Protein  
**Collection Facility:**  
 ESRF beamline ID14-2  
**Data Creation Date:**  
 May 17, 2006  
**Related Datasets:** 99

**Cite this Dataset**  
 Pereira, PJB. 2015. "X-Ray Diffraction data for: *Medicago truncatula* glutamine synthetase. PDB Code 4IS4", SBGrid Data Bank, V1, <http://dx.doi.org/10.15785/SBGRID/100>.

**Dataset Description**  
 Native dataset, low resolution pass

**Reprocessing Instructions**  
 none

**Project Members**

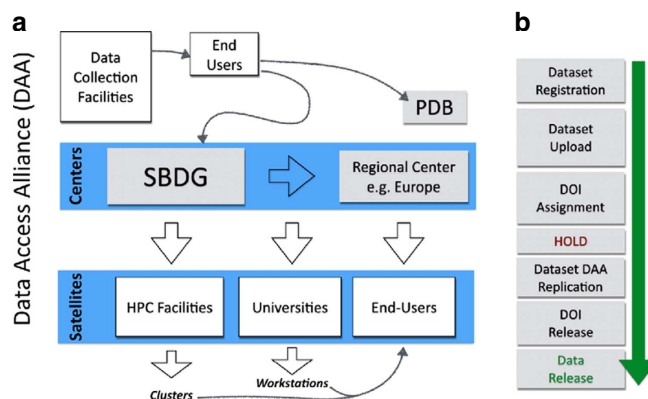
Name	Additional Roles	Affiliation While Working on the Project
Pedro JB Pereira	Data Collector, Depositor	IBMC - Instituto de Biologia Molecular e Celular, Porto, Portugal
Pedro Pereira	PI	Universidade Do Porto

**License and Terms of use**  
 License: [CC0](#)  
 Terms: Our [Community Norms](#) as well as good scientific practices expect that proper credit is given via citation. Please use the data citation, as generated by the SBGrid Data Bank.

**Figure 4 | SBDG persistent data set landing page (the target of a DOI resolver for a published data set).** Data set metadata are displayed, as are instructions for downloading and verifying the data set.

Argonne, IL, replicates all SBDG data sets that originate from NE-CAT beamlines and makes them available to beamline staff and users. Another SBGrid member and DAA Satellite, Yale University, replicates all data sets from Yale laboratories on its institutional storage and makes them accessible to structural biology workstations through the Network File System. We expect that, as research storage infrastructure catches up with the capacities required to archive larger collections of diffraction data sets, some DAA satellites will elect to replicate a larger fraction of SBDG archives and make them available to the general community.

While the DAA offers a variety of data access options that will support growth of the repository, members of the community can also download individual data sets directly from SBGrid servers at Harvard using an rsync protocol. Instructions for downloading



**Figure 5 | Experimental data flow and publication.** (a) Flow of Primary Experimental Data. Data sets collected at synchrotrons are moved to end-users' computers for processing and structure determination. Subsequently refined macromolecular models are deposited at PDB and primary data is uploaded to SBDG. From SBDG, data sets are replicated to DAA centres and eventually copied to DAA Satellites. End-users can access data sets by downloading from DAA centres and by direct access from Satellites. (b) Flowchart for data publication.

individual data sets are provided on the Data set View Pages, and effectively all data sets can be downloaded using the following command: 'rsync -av rsync://data.sbgrid.org/DOI.', where DOI is the digital object identifier for a particular data set. The rsync utility, which is native to Linux and OS X systems, is particularly suitable for downloading large data files and can be restarted in case of interruption. After download, the data integrity of individual data sets can be verified by following instructions on the Data Grid website. With a well defined and permissive CC0 access licence and multiple channels for accessing data (four DAA sites and the rsync download mechanism) our initial infrastructure is well suited to support expansion of the data collection.

**Data publication cycle.** For many SBGrid laboratories, interest in data deposition is driven by a desire to better organize research data and comply with institutional, federal, and project-specific data preservation requirements. During the pilot phase, data deposition privileges were limited to SBGrid member laboratories. With recent funding to further support the project, the Data Grid is now open to the entire structural biology community. Non-SBGrid groups would first need to register with the SBDG to obtain proper deposition credentials.

Wide adoption of data preservation systems is often hindered by the complexities involved in the data deposition process itself. To mitigate this problem, SBDG deposition involves two simple steps: registration and uploading (Fig. 5b). To register a data set, the depositor completes a web form with basic information about the sample, data collection facility, related objects (for example, publication, PDB code), and authorship; this information is mapped to the DataCite schema (Fig. 6). Many details necessary for data set reprocessing—beam center, distance, wavelength, and so on—are automatically included with most data sets in the form of an image header generated by the data collection software at the time of collection, simplifying the registration process. A principal investigator is authorized to sponsor depositions as a recognized member of the community and must approve each deposit. This system allows maximum flexibility when accepting data for deposition, facilitating the upload of complex data sets that otherwise could be challenging to validate. Following registration, a DOI is reserved for the data set and the user is

provided with data transfer instructions. Data deposition is handled by an automated script provided by SBDG and run on the depositor's computer, which uploads the data and checks for data integrity after upload. Upon verification, the primary data are either released in the bi-weekly SBDG release or placed on hold. As with the PDB, release of data placed on hold will coincide with publication.

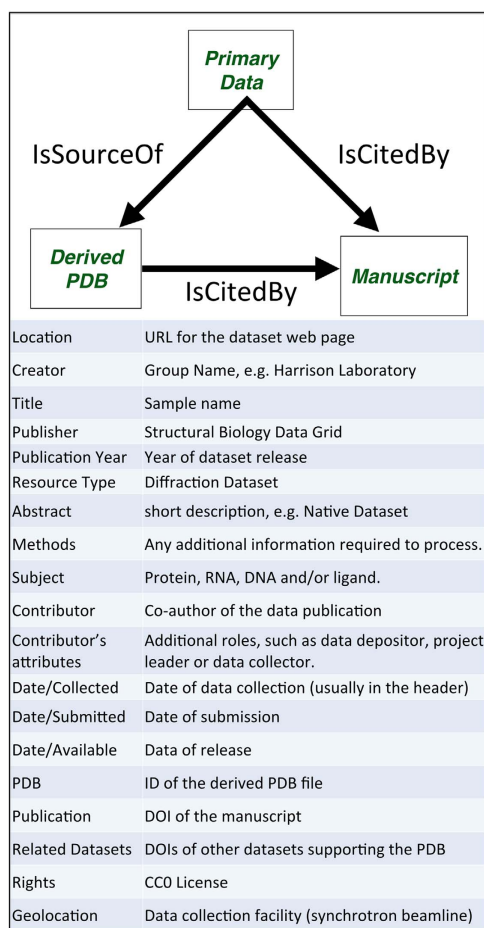
The two-step publication process is complemented by behind-the-scenes data replication, DOI registrations, and data analysis. All X-ray diffraction images are currently post-processed using data processing pipelines that provide a post-publication data review that will be shared with depositors and the community in the next phase of the SBDG project. We are building additional tools to help increase data deposition rates, including automatic reminders sent to consortium members to encourage them to deposit data for previously published work.

**Data citation.** Research data are the legitimate and citable product of research<sup>24,25</sup> 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<sup>24,26,27</sup>. Data citation examples are provided on individual data set pages (Fig. 4). The SBDG complements our AppCiter application<sup>28</sup>, which facilitates citation of research software. Both services are now presented to users in a unified publication support workflow (Fig. 7a). In step 1, the user deposits research-related data that are

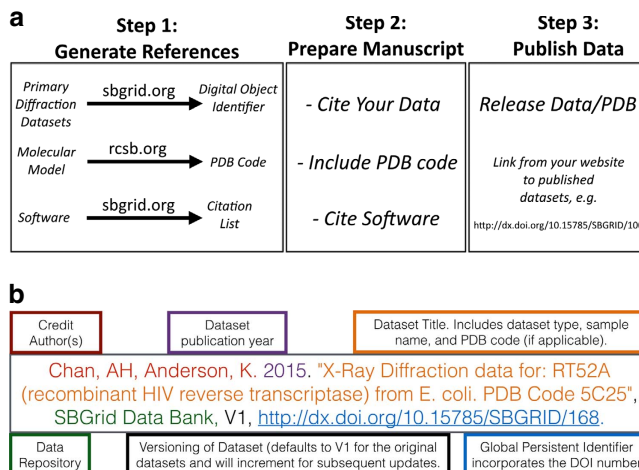
put on hold until publication. A set of DOIs and corresponding data citations are then generated and provided to the end-user. Users can also use AppCiter to generate a list of software citations for all scientific software used in the project. In step 2, all research data and scientific software citations are included in the References section of the manuscript. In step 3 the user, anticipating manuscript publication, contacts relevant databases to request release of the primary and supporting data. This process should, ideally, take place before manuscript publication and be timed to coincide with the publication date, allowing the community to access the data when the manuscript is released. When preparing future publications that refer to completed structures, scientists should reference the relevant publications and macromolecular models, unless they are referring to a specific data set. For specific data sets, authors should explicitly reference experimental data using the corresponding data citation (Fig. 7b). Citation metrics for published data sets will be comparable to those obtained for journal publications.

**Data grid content.** Ease of data deposition and community-wide interest facilitated growth of the initial collection of X-ray diffraction data sets when it opened to the SBGrid community in May 2015. The data sets deposited during the pilot collection phase represent a wide cross-section of structures and a diverse subset of journal articles and structure determination methods. For example, 68 structures derived from data deposited in the SBDG have been determined by molecular replacement, while 4 have been solved by Multiple-wavelength Anomalous Diffraction, 4 by Single Isomorphous Replacement with Anomalous Scattering and 15 by Single-wavelength Anomalous Diffraction. The highest resolution data set<sup>29</sup> extended to 1.04 Å, and the lowest resolution data set<sup>30</sup> to 5.5 Å. The structures ranged in molecular weight from 8.1 (ref. 31) to 426 kDa (ref. 32). The solvent content of these structures ranged from 32 (ref. 33) to 85% (ref. 34) and the longest unit cell edge was reported to be 525.29 Å (ref. 35).

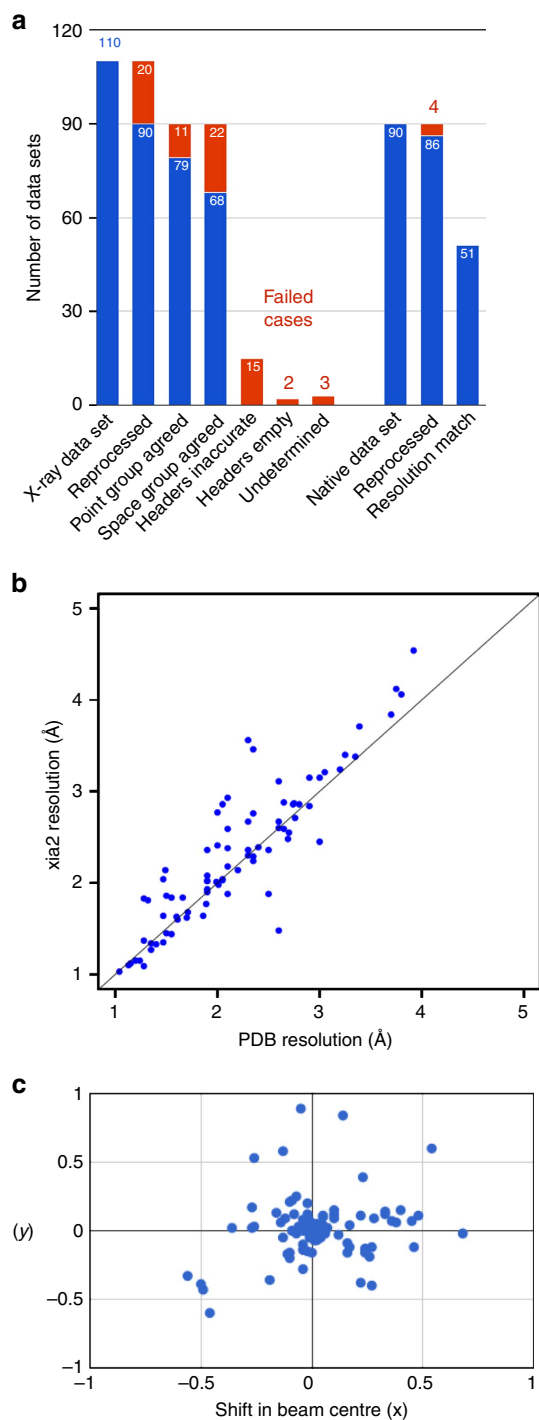
For a proof of concept, released data sets in the SBGridDB were reprocessed with *xia2* (refs 36–42) in a fully automated manner (Fig. 8a). In all, 90 of the 110 released data sets with a corresponding PDB ID were successfully reprocessed. In all, 86 of those 90 data sets represented high-resolution, native data and for 51 of those *xia2* decision making determined a high-resolution limit within 0.1 Å of the published structure (Fig. 8b). The point



**Figure 6 | DataCite metadata schema used for primary data sets within the SBDG.** Information associated with the DOI record for a primary data set through the EZID system.



**Figure 7 | Data publication guidelines.** (a) Flowchart illustrating publication guidelines incorporating software and data citations. (b) Data Citation guidelines, adapted from Dataverse Best Practices Guidelines that were developed based on Force 11 Joint Declaration of Data Citation Principles.



**Figure 8 | Reprocessing of X-ray diffraction data sets.** (a) Analysis of 110 X-ray diffraction data sets that supported previously published PDB coordinates. Most of the failures (represented in red) were due to inaccurate or incomplete image-header information. In several of these cases, depositors provided annotations correcting this information; (b) Comparison of resolution determined by automated *xia2* reprocessing with published resolution. Includes data sets not used for final refinement of published structures; (c) Shift in direct beam position from image headers and refined value following successful reprocessing with *xia2*.

group determined by reprocessing agreed with that of the published structure in 79 cases; for 65 of these the space groups agreed. The lower degree of recovery of space groups, in comparison to point groups, is attributed to ambiguity in screw

axis determination at this stage of data processing. To provide insight into the most common failure modes, data sets for which *xia2* did not produce a set of integrated intensities were investigated using iMOSFLM<sup>43</sup>. Twelve of the failure cases could be attributed to absent or inaccurate information in the image headers: while accuracy of the beam center annotation varied within the pilot collection (Fig. 8c), 10 data sets had visually incorrect beam center information, two had missing header information. The cause of failure for the eight remaining data sets was not definitively determined from the data sets alone; however, consulting the reprocessing instructions provided by depositors clarified this for five of these data sets. The reprocessing instructions also suggested that many of the data sets for which *xia2* was able to produce integrated intensities, but with resolution or symmetry disagreeing with the deposited structure, could be attributed to incorrect header information. One outlying reprocessing case for which a significantly higher resolution was determined than originally reported was also investigated. For this case, one of four reprocessing attempts for the data set reported a resolution higher than that supported by merging statistics. This discrepancy was resolved by a software update.

In addition to estimates of the Bragg intensities, diffraction images can also be analysed for additional features<sup>44</sup>. A well-known example is the isotropic solvent ring that generally appears  $\sim 3\text{--}4\text{ \AA}$  resolution<sup>45</sup>. However, diffraction images also contain anisotropic diffuse scattering signals under and between the Bragg peaks that derive from two-point correlations of electron density fluctuations<sup>7</sup>. Analysis of this diffuse scattering could therefore provide information about protein, nucleic acid, and lipid structural dynamics and correlated motions, potentially leading to new mechanistic insights<sup>46</sup> or to validating sampling schemes and energy functions for molecular dynamics simulations<sup>47</sup>. One data set on the model enzyme Cyclophilin A is currently deposited (Table 2) to be used as ‘gold-standard’ to compare the influence of temperature on data collection<sup>48</sup> and to assess consistency between X-Ray Free-Electron Laser (XFEL) and synchrotron data<sup>49</sup>. This data set can now also be analysed for diffuse scattering features, which could distinguish between models of correlated motion suggested by NMR experiments.

**X-ray diffraction reference subset and other collections.** To take advantage of data grid diversity, we have selected a small subset of cases that could be used to support software development and teaching of data processing and diverse structure determination techniques (Table 2). This subset includes high-resolution (1.2 Å), low resolution (4.5 and 7.0 Å), anisotropic and twinned data sets. Additionally data sets that supported a variety of experimental phasing approaches (for example, phasing with selenium, zinc, uranium, barium/potassium) and molecular replacement cases (for example, with a 9 Å electron microscopy (EM) envelope) are included. The subset also incorporates diffraction data for crystals grown in lipidic cubic phase and an example of multi-crystal averaging.

Additionally SBDG is suited to support various other primary data types that are being generated by members of the consortium, and those pilot collections will seed development of community-wide data analysis systems. MicroED is a promising new technique<sup>50,51</sup> and inclusions of the early microcrystal data sets might stimulate the community to explore this technique and to fine-tune data processing software. Examples of MicroED data sets that are included in the pilot collection include three MicroED data sets that were used to determine structures of the toxic core of  $\alpha$ -synuclein<sup>52</sup>, catalase<sup>53</sup> and lysozyme<sup>54</sup>. Other types of data sets in our pilot collection include a 55 GB computational decoy data set for

**Table 2 | Reference subset.**

Data set	Description
10.15785/SBGRID/5 Boggon Laboratory Reference Case 1: MR/Multi-crystal averaging.	Data sets from 5 crystals of SNX17 FERM domain in complex with a peptide corresponding to KRIT1's NPXY2 motif. Separate integration of the data sets and scaling together allows a complete 3.0 Å data set for molecular replacement solution (original paper used 4GXB as a search model) and structure refinement.
10.15785/SBGRID/117 Baxter Laboratory Reference Case 2: MR/Low resolution, twinned with rotational pseudosymmetry.	3.70 Å data set collected on a crystal of thioester-containing protein 1 *S1 allele (TEP1*S1). Initial data processing suggested $P4_32_12$ , but one of the two molecules (~1300 aa. each) in the ASU overlapped with its symmetry-mate. Comparison of alternative scenarios in refinement identified the true space group as $P4_3$ with twinning and rotational pseudosymmetry. Refinement was completed with TLS, NCS (local) and external restraints derived by <i>ProSMART</i> <sup>65</sup> using TEP1*R1 (PDB 4D94) as reference.
10.15785/SBGRID/62 Modis Laboratory Reference case 3: U SAD/Low resolution.	4.5 Å data set 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 Laboratory Reference Case 4: Ba/K SAD; 91 nt RNA-chromophore complex.	2.5 Å data set collected at ALS BL 5.0.2 using 6.0 keV X-rays from a crystal of 'Spinach' a fluorescent RNA analogue 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 Reference Case 5: Zn SAD; 4 Zn/ASU protein/RNA complex.	2.9 Å Zn SAD data set 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 Reference Case 6: 3.29-Å SeMet SAD 9 Se/ASU	This 3.29-Å selenomethionine SAD data set, collected at 0.9789 Å wavelength at BNL X25 beamline, was sufficient to determine the phases and to trace the structure of HSV-2 gH/gL complex <sup>66</sup> . There are 9 Se sites in the ASU. During integration in HKL2000, $\chi^2$ 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 ( $a = b = 88$ Å versus $c = 333$ Å).
10.15785.SBGRID/179 Schwartz Laboratory Reference Case 7: MR-SAD at 7.0 Å	Contaminating <i>E.coli</i> protein 4FCC_A, acting as a crystallization chaperone, was found readily by MR. Using these MR phases seven $(Ta_6Br_{12})^{2+}$ -positions could be found in the 8.8 Å derivative data set 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/218 10.15785/SBGRID/78 Rudenko Laboratory Reference Case 8: MR-SAD at 2.65 Å (44 Se atoms/ASU) 10.15785/SBGRID/9 Tao Laboratory Reference case 9: 3.25 Å data set used for MR with a 9-Å cryo-EM envelope	3.25 Å data set (#218) from a crystal of the selenomethionyl neurexin 1α ectodomain and 2.65 Å higher resolution native data set (#78), both collected at APS using multiple settings. The structure has 2 molecules/ASU with a total of 14 ordered domains and ~2,000 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 the 3.25 Å SeMet SAD data set. Refinement was completed using data set #78. A 3.25-Å resolution data set 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.
10.15785/SBGRID/83 Drennan Laboratory Reference Case 10: MR/large unit cell, anisotropic.	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 data set that can be used for molecular replacement.
10.15785/SBGRID/125 Kruse Laboratory (data collected in Kobilka Laboratory) Reference Case 11: MR, lipidic cubic phase	Diffraction data for lipidic cubic phase crystals of human M <sub>2</sub> muscarinic acetylcholine receptor bound to the agonist iperoxo, the allosteric modulator LY2119620, and the conformationally-selective nanobody Nb9-8.
DOI: 10.15785/SBGRID/68 Fraser Laboratory Reference case 12: X-ray diffuse scattering	1.2 Å data set collected at SSRL provides a high-resolution standard data set of the enzyme Cyclophilin to examine the influence of data collection temperature to compare with XFEL data, and to measure X-ray diffuse scattering.

MR, molecular replacement; SAD, Single-wavelength Anomalous Diffraction.

12 X-ray diffraction data sets from the SBDG pilot collection were identified as particularly suitable for software testing and teaching activities. In addition, data sets from molecular dynamics, lattice light-sheet microscopy and MicroED represent an invaluable subset.

55 complexes with associated HADDOCK scores<sup>55</sup>, a 2  $\mu$ s Desmond<sup>56</sup> MD trajectory<sup>57</sup>, and a recently collected Lattice Light-Sheet Microscopy<sup>58,59</sup> data set with in-vivo imaging of zebrafish embryos<sup>60</sup>. Here the engagement with domain experts and respective communities will be also required to establish data validation pipelines and effective DAA distribution models.

## Discussion

We have developed a flexible data publication system, the SBDG, to support deposition of a variety of large primary data sets. The data repository complements the wwPDB efforts by preserving the raw data that supports PDB-deposited structure models. The pilot phase of the project, which was limited to SBGrid laboratories, demonstrated both feasibility and strong participation, with the deposition and publication of 117 data sets (as of 1 September 2015, collected over 3 months). To support annotated data collection, we have established data processing pipelines that will evolve the post-deposition data-analysis process. For example, the pipeline presented in the results section allows depositors and SBDG curators to quickly identify image-header problems, and parameters that are refined or corrected will be included in the expanded Dataverse schema<sup>61–64</sup>. The outliers and failures of the current reprocessing pipeline illustrate areas of potential improvement to metadata accuracy and the pipeline itself. Data depositors and other community members will be able to provide data annotations to assist with the convergence of this process. Access to this growing collection of X-ray diffraction data sets will support the proposed paradigm shift in the community<sup>6</sup> from the static archive towards a much more dynamic body of continuously improving refined models.

Despite being in the age of ‘big data science’, universal storage of large, biomedical data sets is an issue that has not yet been resolved, as infrastructure and support responsibilities have not been well defined. Shifting the burdens of data management from individual research groups and institutions to global infrastructures is an effective and economical strategy to address this issue that has previously been proven successful by the wwPDB and would now be demonstrated by the SBDG. By virtue of the consortium’s global presence, SBDG is well positioned to stimulate community-wide participation. SBGrid may facilitate integration of the data grid with regional projects and facility-related efforts to preserve primary diffraction data sets. This data distribution model is similar to those established in other fields. For example, the Data Preservation Alliance ([www.data-pass.org](http://www.data-pass.org)) replicates and indexes quantitative data for the social sciences. Data collected at the Large Hadron Collider are made available under a multi-tier processing and storage framework. As a large international consortium backed by diverse funding mechanisms and DAA storage contributions of its members, SBGrid is uniquely capable of bypassing grant limitations that would otherwise deter such a long-term global infrastructure effort. Given recently secured funding to support data curation and technology integration under the Dataverse research data management, and with gradual community investment, SBDG is poised to scale up to support the entire community.

While access to experimental data is critical to ensuring research reproducibility, metadata quality is also crucial. Data sets that are poorly annotated have limited use to the research community. With a focus on deployment of a sustainable and flexible data management infrastructure, the SBDG takes a unique approach on metadata preservation. The repository employs an accommodating DataCite schema, which preserves basic information about experiments. The depositions are self-moderated by contributing laboratories, with data publication subject to approval of the principal investigators. As our results

demonstrated, this approach worked well for the vast majority of data sets deposited in the SBDG, 82% of which were automatically reprocessed with current data processing software and the majority of the remaining data sets could be easily reprocessed manually. This success rate for reprocessing diffraction data sets was achieved without any explicit quality control to ensure that the data sets contained sufficient information for reprocessing—in other words, using image headers as the only source of experimental (geometry and detector) parameters. Two possibilities under consideration for maintaining and improving this success rate are allowing depositors to annotate updated experimental parameters (for example, beam center) and explicit checks for metadata required for reprocessing prior to data publication. To facilitate interoperability with other projects and further stimulate uniform data evaluations, we will work in parallel to develop tools that will support download of archived data sets in community accepted master formats supporting intrinsic metadata, such as OME-TIFF or HDF5. This process will allow annotation of downloadable data sets with additional information from analysis pipelines, and will be guided by feedback from projects that interface with SBDG. Ideally, publication of data sets will encourage the communities to adopt standardized formats and ensure complete population of experimental metadata with adequate accuracy to support reprocessing.

While the SBDG immediately serves the well-defined area of X-ray crystallography, our pilot project has demonstrated that our infrastructure can preserve additional data types, such as decoy data sets for NMR computations or MicroED data sets. SBDG will duplicate XFEL data sets that are currently accessible through the Coherent X-ray Imaging Data Bank (<http://www.cxidb.org/>) and support their distribution by DAA. In addition, SBDG will collaborate with MicroED and XFEL collection curators who will moderate development of community driven efforts to automate data analysis pipelines to parallel automatic processing of X-ray diffraction data sets with packages like DIALS or *xia2*. We envision that the tools and technologies that arise from this project will ultimately lead to the development of a fully featured, primary data publication system. Features of such a system would include the capability of supporting a variety of experimental data types and automatic incorporation of pertinent data set information during data collection at local, regional and national facilities. The integration of primary data management with a base set of scientific software enables repositories to progress towards dynamically improving sources of knowledge, as well as providing an integrated computing environment for ongoing research.

In summary, we have presented the SBDG, a new system for the preservation and publication of large experimental data sets. The system is the latest product of SBGrid’s mission to maintain a community-wide research-software infrastructure. Through disclosure, adoption, transparency, management of external dependencies, permissible licensing, and technical protection mechanisms, the SBDG is committed to compliance with evolving community standards of data preservation. We expect that the widespread sharing of experimental data will support methods development and will ultimately lead to better quality of structural models that are subject to continuous methods improvement.

## Methods

**Current implementation.** The databank deposition process involves five stages: (1) recording associated metadata, (2) local checksum calculation, (3) data transfer, (4) post-transfer verification and (5) public identifier registration.

A publicly accessible web frontend is used for handling user interactions with the databank. Built using the Python-based Django web framework, this frontend runs on an Ubuntu 14 LTS Server with a PostgreSQL 9.3 database. It collects the necessary metadata during deposition and informs the backend systems about deposition requests. A cryptographic checksum (SHA1 SUM, FIPS 180-4) is calculated before

data transfer. This ensures that the data set is unchanged. Data transfer is handled by rsync over ssh. Once data transfer is complete, the databank verifies that the data set has been transferred uncorrupted, or reports a problem with the data set. If necessary, extraneous files (intermediate data files, processing or transfer scripts) are removed, data files are uncompressed and checksums re-computed. In the event of any modifications to the data set, an unmodified copy is stored in an offline file system. Upon data set release, the DOI reserved during data set registration is registered using the recorded metadata, and the data set (including checksum information) is made available for download over anonymous rsync.

**Metadata schema.** DOIs are issued through EZID, through the Harvard University Library and the California Digital Library. Metadata are organized following the DataCite schema (Fig. 6).

**Reprocessing details.** Data sets that had been publicly released by 1 September 2015 were reprocessed by *xia2* in a fully automated manner. For each data set, four attempts were made to reprocess using options '-2d', '-3d', '-3dii' and '-dials', using MOSFLM, XDS, XDS (indexing with peaks from all images) and DIALS, respectively. AIMLESS<sup>37</sup> and POINTLESS<sup>38</sup> were used by *xia2* for space group determination. A data set was considered successfully reprocessed if any of these attempts succeeded, and comparisons to the originally published structure were done with the best matching result. Investigation of unsuccessfully reprocessed data sets was performed using iMOSFLM<sup>41</sup>. This investigation was performed 'blinded' to the reprocessing instructions provided by depositors, in order to better investigate the limits of relying solely on diffraction images.

**Data alliance.** Released data sets are distributed to Data Alliance mirror sites using the same mechanism as individual data set distribution. Data set checksums enable accurate data transfer. Users can select a mirror site by picking an appropriate rsync URL for data download.

## References

- Bilderback, D. H., Elleaume, P. & Weckert, E. Review of third and next generation synchrotron light sources. *J. Phys. B: At. Mol. Opt. Phys.* **38**, S773–S797 (2005).
- Guss, J. M. & McMahon, B. How to make deposition of images a reality. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2520–2532 (2014).
- Meyer, G. R. *et al.* Operation of the Australian Synchrotron for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2510–2519 (2014).
- Elslinger, M.-A. *et al.* The JCSG high-throughput structural biology pipeline. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1137–1142 (2010).
- Kroon-Batenburg, L. M. J. & Helliwell, J. R. Experiences with making diffraction image data available: what metadata do we need to archive? *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2502–2509 (2014).
- Terwilliger, T. C. & Bricogne, G. Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2533–2543 (2014).
- Wall, M. E., Adams, P. D., Fraser, J. S. & Sauter, N. K. Diffuse X-ray scattering to model protein motions. *Structure* **22**, 182–184 (2014).
- Joosten, R. P. *et al.* PDB REDO: automated re-refinement of X-ray structure models in the PDB. *J. Appl. Crystallogr.* **42**, 376–384 (2009).
- Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).
- Tanley, S. W. M., Diederichs, K., Kroon-Batenburg, L. M. J., Schreurs, A. M. M. & Helliwell, J. R. Experiences with archived raw diffraction images data: capturing cisplatin after chemical conversion of carboplatin in high salt conditions for a protein crystal. *J. Synchrotron Radiat.* **20**, 880–883 (2013).
- Matthews, B. W. Five retracted structure reports: inverted or incorrect? *Protein Sci.* **16**, 1013–1016 (2007).
- Janssen, B. J., Read, R. J., Brunger, A. T. & Gros, P. Crystallography: crystallographic evidence for deviating C3b structure. *Nature* **448**, E1–E2 (2007).
- Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Biol.* **10**, 980–980 (2003).
- Berman, H., Kleywegt, G., Nakamura, H. & Markley, J. The protein data bank archive as an open data resource. *J. Comput. Aided Mol. Des.* **28**, 1009–1014 (2014).
- Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
- Foster, I. in *Network and Parallel Computing* (eds Jin, H., Reed, D. & Jiang, W.) volume 3779, 2–13 (Springer, 2005).
- Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing* **15**, 70–73 (2011).
- Chard, K. *et al.* in 2015 IEEE 11th International Conference on e-Science (e-Science), 401–410 (Munich, 2015).
- Stokes-Rees, I. *et al.* Adapting federated cyberinfrastructure for shared data collection facilities in structural biology. *J. Synchrotron Radiat.* **19**, 462–467 (2012).
- Lee, D. & Raman, C. X-Ray Diffraction data for: *Escherichia coli* DOS Br complex. PDB Code 1V9Z, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/137> (2015).
- Rudenko, G. X-Ray Diffraction data for: neurexin 1alpha extracellular domain. PDB Code 3QCW, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/78> (2015).
- Biasini, M. PV-WebGL-based protein viewer. *Zenodo* doi:10.5281/zenodo.12620 (2014).
- Starr, J. *et al.* Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput. Sci.* **1**, e1 (2015).
- Martone, M. (ed). Data citation synthesis group: Joint declaration of data citation principles. FORCE11; <https://www.force11.org/datacitation> (2014).
- Bourne, P. E. *et al.* Improving the future of research communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). *Dagstuhl Manifestos* **1**, 41–60 (2011).
- Altman, M. & King, G. A proposed standard for the scholarly citation of quantitative data. *D-lib Mag.* **13** (2007).
- Altman, M. & Crosas, M. The evolution of data citation: from principles to implementation. *IASSIST Q.* **37**, 62 (2013).
- Socias, S., Morin, A., Timony, M. & Sliz, P. AppCiter: a web application for increasing rates and accuracy of scientific software citation. *Structure* **23**, 807–808 (2015).
- Hunter, J. C. & Westover, K. D. X-Ray Diffraction data for: Human GT- Pase KRAS G12R bound to GDP. PDB Code 4QL3, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/160> (2015).
- Gilman, M. S. A. & McLellan, J. S. X-Ray diffraction data for: Motavizumab and AM14 in complex with prefusion RSV f. PDB code 4ZYP, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/155> (2015).
- Feldkamp, M. D. & Chazin, W. J. X-Ray diffraction data for: human RPA32C. PDB code 4OU0, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/92> (2015).
- Tolia, N. H. X-Ray diffraction data for: erythrocyte binding antigen 140. PDB code 4GF2, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/115> (2015).
- Hunter, J. C. & Westover, K. D. X-Ray diffraction data for: human GTPase KRAS G12C bound to GDP. PDB code 4LDJ, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/158> (2015).
- Corbett, K. D. & Harrison, S. X-Ray diffraction data for: *S. cerevisiae* Csm1-Mam1 complex. PDB code 4EMC, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/24> (2015).
- Gajadeera, C. S. & Tsodikov, O. V. X-Ray diffraction data for: Inorganic pyrophosphatase from staphylococcus aureus in complex with mn2+. PDB code 4RPA, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/22> (2015).
- Winter, G., Loble, C. M. C. & Prince, S. M. Decision making in xia2. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1260–1273 (2013).
- Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013).
- Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 72–82 (2006).
- Kabsch, W. XDS. *Acta. Cryst.* **66**, 125–132 (2010).
- Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
- Leslie, A. G. W. Integration of macromolecular diffraction data. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1696–1702 (1999).
- Waterman, D. G. *et al.* The DIALS framework for integration software. *CCP4 Newsl. Protein Crystallogr.* **49**, 13–15 (2013).
- Battye, G. T., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 271–281 (2011).
- Helliwell, J. R. & Mitchell, E. P. Synchrotron radiation macromolecular crystallography: science and spin-offs. *IUCrJ* **2**, 283–291 (2015).
- Welberry, T. *Diffuse X-Ray Scattering and Models of Disorder* (OUP Oxford, 2004).
- Wall, M. E., Clarage, J. B. & Phillips, Jr G. N. Motions of calmodulin characterized using both bragg and diffuse X-ray scattering. *Structure* **5**, 1599–1612 (1997).
- Wall, M. E. *et al.* Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering. *Proc. Natl. Acad. Sci.* **111**, 17887–17892 (2014).
- Wall, M. Methods and software for diffuse X-ray scattering from protein crystals. In *Micro and Nano Technologies. In Bioanalysis Methods in Molecular Biology* (eds Foote, R. S. & Lee, J. W.) volume 544, 269–279 (Humana Press, 2009).

49. Fraser, J. S. X-Ray diffraction data for: Cyclophilin a. PDB code 4YUO, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/68> (2015).
50. Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. Three-dimensional electron crystallography of protein microcrystals. *Elife* **2**, e01345 (2013).
51. Nannenga, B. L., Shi, D., Leslie, A. G. & Gonen, T. High-resolution structure determination by continuous-rotation data collection in MicroED. *Nat. Methods* **11**, 927–930 (2014).
52. Reyes, F., Rodriguez, J. & Gonen, T. Micro-Electron diffraction data for: alpha-synuclein. PDB code 4RIL, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/193> (2015).
53. de la Cruz, J., Shi, D. & Gonen, T. Micro-Electron diffraction data for: bovine catalase. PDB code 3J7B, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/186> (2015).
54. Shi, D. & Gonen, T. Micro-Electron diffraction data for: Hen egg white lysozyme. PDB code 3J6K, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/185> (2015).
55. Vangone, A. & Bonvin, A. M. HADDOCK docking models, Structural Biology Data Grid, volume V1, <http://dx.doi.org/10.15785/SBGRID/131> (2015).
56. Bowers, K. *et al.* in *Proceedings of the ACM/IEEE SC 2006 Conference*, 43–43 (Tampa, FL, USA, 2006).
57. Sliz, P. Molecular dynamics trajectory of human O-GlcNAc transferase. PDB code 3PE4, Structural Biology Data Grid, <http://dx.doi.org/10.15785/SBGRID/190> (2015).
58. Chen, B.-C. *et al.* Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science* **346**, 1257998 (2014).
59. Kural, C. *et al.* Asymmetric formation of coated pits on dorsal and ventral surfaces at the leading edges of motile cells and on protrusions of immobile cells. *Mol. Biol. Cell* **26**, 2044–2053 (2015).
60. Upadhyayula, S. & Kirchhausen, T. Lattice Light-Sheet microscopy data for: developing zebrafish embryo, Structural Biology Data Grid, V1, <http://dx.doi.org/10.15785/SBGRID/187> (2015).
61. Crosas, M., Honaker, J., King, G. & Sweeney, L. Automating open science for big data. *Ann. Am. Acad. Polit. Soc. Sci.* **659**, 260–273 (2015).
62. Crosas, M. A data sharing story. *eScience Librariansh.* **1**, 7 (2013).
63. Crosas, M. The dataverse network: an open-source application for sharing, discovering and preserving data. *D-lib Mag.* **17**, 2 (2011).
64. King, G. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociol. Meth. Res.* **36**, 173–199 (2007).
65. Nicholls, R. A., Fischer, M., Stuart, M. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2487–2499 (2014).
66. Chowdary, T. K. *et al.* Crystal structure of the conserved herpesvirus fusion regulator complex gH-gL. *Nat. Struct. Mol. Biol.* **17**, 882–888 (2010).

## Acknowledgements

Development of the Structural Biology Data Grid is funded by The Leona M. and Harry B. Helmsley Charitable Trust 2016PG-BRI002 to PS and MC. Development of citation workflows is supported NSF 1448069 (to PS). DAA is being developed as a pilot project of the National Data Service, with additional funds to support storage and technology development, including NIH P41 GM103403 (NE-CAT) and 1S10RR028832 (HMS) and DOE DE-AC02-06CH11357; NIH 1U54EB020406-01, Big Data for Discovery Science Center; and NIST 60NANB15D077 (Globus Project). AB acknowledges Ariel Chaparro for assistance with the DAA setup (Inst Pasteur Montevideo). Collections of pilot data sets were supported by various grants (see Supplementary Table 1).

## Author contributions

All authors contributed to the current study, including intellectual input and editing of the manuscript. P.A.M., S.S. and P.S. developed the data grid system and A.B., M.L., C.B., J.W., D.N., K.R., J.K., F.R.N.C.M., I.F., MC and P.S. implemented the Data Access Alliance infrastructure. P.A.M., K.D. and P.S. analysed the data. K.S.A., R.H.B., S.C.B., T.J.B., D.B., T.J.B., A.C., C.I.C., W.J.C., K.D.C., M.S.C., S.C., S.D.P., E.D.C., C.L.D., M.J.E., B.F.E., Q.R.F., A.R.F., J.S.F., J.C.F., K.C.G., R.G., P.G., S.C.H., E.E.H., Z.J., R.J.K., A.C.K., M.K., J.S.M., Y.M., Y.N., Z.O., E.F.P., P.J.B.P., C.P., C.S.R., T.A.R., A.R., M.K.R., G.R., J.S., T.U.S., Y.S., H.S., Y.J.T., N.H.T., O.V.T., K.D.W., H.W. and P.S. contributed X-ray diffraction data sets. A.M.J.J.B contributed the HADDOCK docking decoys data set. T.G., T.K., and P.S. contributed MicroED, Lattice Light-Sheet Microscopy and Molecular Dynamics data sets, respectively. P.A.M., E.R. and P.S. Analysed the data and wrote the paper.

## Additional information

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions>.

**How to cite this article:** Meyer, P. A. *et al.* Data publication with the structural biology data grid supports live analysis. *Nat. Commun.* **7**:10882 doi: 10.1038/ncomms10882 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Peter A. Meyer<sup>1</sup>, Stephanie Socias<sup>1</sup>, Jason Key<sup>1</sup>, Elizabeth Ransey<sup>1</sup>, Emily C. Tjon<sup>1</sup>, Alejandro Buschiazzi<sup>2,3</sup>, Ming Lei<sup>4</sup>, Chris Botka<sup>5</sup>, James Withrow<sup>6</sup>, David Neau<sup>6</sup>, Kanagalaghatta Rajashankar<sup>6</sup>, Karen S. Anderson<sup>7</sup>, Richard H. Baxter<sup>8</sup>, Stephen C. Blacklow<sup>1</sup>, Titus J. Boggon<sup>7</sup>, Alexandre M.J.J. Bonvin<sup>9</sup>, Dominika Borek<sup>10</sup>, Tom J. Brett<sup>11</sup>, Amedeo Caffisch<sup>12</sup>, Chung-I Chang<sup>13</sup>, Walter J. Chazin<sup>14</sup>, Kevin D. Corbett<sup>15,16</sup>, Michael S. Cosgrove<sup>17</sup>, Sean Crosson<sup>18</sup>, Sirano Dhe-Paganon<sup>19</sup>, Enrico Di Cera<sup>20</sup>, Catherine L. Drennan<sup>21</sup>, Michael J. Eck<sup>1,19</sup>, Brandt F. Eichman<sup>22</sup>, Qing R. Fan<sup>23</sup>, Adrian R. Ferré-D'Amaré<sup>24</sup>, J. Christopher Fromme<sup>25</sup>, K. Christopher Garcia<sup>26,27,28</sup>, Rachele Gaudet<sup>29</sup>, Peng Gong<sup>30</sup>, Stephen C. Harrison<sup>1,31,32</sup>, Ekaterina E. Heldwein<sup>33</sup>, Zongchao Jia<sup>34</sup>, Robert J. Keenan<sup>18</sup>, Andrew C. Kruse<sup>1</sup>, Marc Kvsanskul<sup>35</sup>, Jason S. McLellan<sup>36</sup>, Yorgo Modis<sup>37</sup>, Yunsun Nam<sup>38</sup>, Zbyszek Otwinowski<sup>10</sup>, Emil F. Pai<sup>39,40</sup>, Pedro José Barbosa Pereira<sup>41</sup>, Carlo Petosa<sup>42</sup>, C.S. Raman<sup>43</sup>, Tom A. Rapoport<sup>44</sup>, Antonina Roll-Mecak<sup>45,46</sup>, Michael K. Rosen<sup>47</sup>, Gabby Rudenko<sup>48</sup>, Joseph Schlessinger<sup>49</sup>, Thomas U. Schwartz<sup>50</sup>, Yousif Shamoo<sup>51</sup>, Holger Sondermann<sup>52</sup>, Yizhi J. Tao<sup>51</sup>, Niraj H. Tolia<sup>53</sup>, Oleg V. Tsodikov<sup>54</sup>, Kenneth D. Westover<sup>55</sup>, Hao Wu<sup>1,56</sup>, Ian Foster<sup>57</sup>, James S. Fraser<sup>58</sup>, Filipe R.N.C. Maia<sup>59,60</sup>, Tamir Gonen<sup>61</sup>, Tom Kirchhausen<sup>62,63</sup>, Kay Diederichs<sup>64</sup>, Mercè Crosas<sup>65</sup> & Piotr Sliz<sup>1</sup>

<sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>2</sup>Laboratory of Molecular & Structural Microbiology, Institut Pasteur de Montevideo, Montevideo 11400, Uruguay. <sup>3</sup>Department of Structural Biology & Chemistry, Institut Pasteur, 75015

Paris, France. <sup>4</sup> Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China. <sup>5</sup> Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>6</sup> NE-CAT and Department of Chemistry and Chemical Biology, Cornell University, Building 436E, Argonne National Laboratory, 9700S. Cass Avenue, Argonne, Illinois 60439, USA. <sup>7</sup> Departments of Pharmacology and Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, Connecticut 06520, USA. <sup>8</sup> Department of Chemistry, Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. <sup>9</sup> Bijvoet Center, Faculty of Science, Utrecht University, 3584 CH Utrecht, The Netherlands. <sup>10</sup> Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. <sup>11</sup> Department of Internal Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>12</sup> Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland. <sup>13</sup> Institute of Biological Chemistry, Academia Sinica, Taipei 11529, Taiwan. <sup>14</sup> Departments of Biochemistry and Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, USA. <sup>15</sup> Ludwig Institute for Cancer Research, San Diego Branch, La Jolla, California 92093, USA. <sup>16</sup> Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA. <sup>17</sup> Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University, Syracuse, New York 13210, USA. <sup>18</sup> Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA. <sup>19</sup> Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. <sup>20</sup> Edward A. Doisy Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, St Louis, Missouri 63104, USA. <sup>21</sup> Departments of Chemistry and Biology and the Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>22</sup> Department of Biological Sciences and Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, USA. <sup>23</sup> Departments of Pharmacology and Pathology and Cell Biology, Columbia University, New York, New York 10032, USA. <sup>24</sup> Laboratory of RNA Biophysics, National Heart, Lung and Blood Institute, NIH, Bethesda, Maryland 20892, USA. <sup>25</sup> Department of Molecular Biology and Genetics, Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York 14853, USA. <sup>26</sup> Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA. <sup>27</sup> Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. <sup>28</sup> Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA. <sup>29</sup> Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>30</sup> Key Laboratory of Special Pathogens and Biosafety, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China. <sup>31</sup> Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>32</sup> Laboratory of Molecular Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>33</sup> Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, Massachusetts 02111, USA. <sup>34</sup> Department of Biomedical and Molecular Sciences, Queen's University, Kingston, Ontario, Canada K7M 3G5. <sup>35</sup> Department of Biochemistry and Genetics, La Trobe University, Melbourne, Victoria 3086, Australia. <sup>36</sup> Department of Biochemistry, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA. <sup>37</sup> Department of Medicine, University of Cambridge, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK. <sup>38</sup> University of Texas, Southwestern Medical Center, Dallas, Texas 75390, USA. <sup>39</sup> Departments of Biochemistry, Medical Biophysics and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8. <sup>40</sup> Campbell Family Institute for Cancer Research, Ontario Cancer Institute/University Health Network, Toronto, Ontario, Canada M5G 2M9. <sup>41</sup> IBMC—Instituto de Biologia Molecular e Celular and Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4150 Porto, Portugal. <sup>42</sup> Université Grenoble Alpes/CNRS/CEA, Institut de Biologie Structurale, 38027 Grenoble, France. <sup>43</sup> Department of Pharmaceutical Sciences, University of Maryland, Baltimore, Maryland 21201, USA. <sup>44</sup> Howard Hughes Medical Institute and Harvard Medical School, Department of Cell Biology, Boston, Massachusetts 02115, USA. <sup>45</sup> Cell Biology and Biophysics Unit, Porter Neuroscience Research Center, National Institute of Neurological Disorders and Stroke, Bethesda, Maryland 20892, USA. <sup>46</sup> National Heart, Lung and Blood Institute, Bethesda, Maryland 20892, USA. <sup>47</sup> Department of Biophysics and Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. <sup>48</sup> Department of Pharmacology and Toxicology, Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, Texas 77555, USA. <sup>49</sup> Department of Pharmacology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. <sup>50</sup> Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>51</sup> Department of BioSciences, Rice University, Houston, Texas 77005, USA. <sup>52</sup> Department of Molecular Medicine, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA. <sup>53</sup> Department of Molecular Microbiology, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>54</sup> Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, Lexington, Kentucky 40536, USA. <sup>55</sup> Departments of Biochemistry and Radiation Oncology, University of Texas, Southwestern Medical Center, Dallas, Texas 75390, USA. <sup>56</sup> Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, Massachusetts 02115, USA. <sup>57</sup> Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, and Department of Computer Science, University of Chicago, Chicago, Illinois 60637, USA. <sup>58</sup> Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA. <sup>59</sup> Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden. <sup>60</sup> NERSC, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. <sup>61</sup> Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147 USA. <sup>62</sup> Program in Cellular and Molecular Medicine and Department of Pediatrics, Boston Children's Hospital, Boston, Massachusetts 02115, USA. <sup>63</sup> Departments of Cell Biology and Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>64</sup> Department of Biology, University of Konstanz, D-78457 Konstanz, Germany. <sup>65</sup> Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138, USA