STRUCTURAL
BIOLOGY

ISSN 2059-7983

The *XChemExplorer* graphical workflow tool for routine or large-scale protein–ligand structure determination

Tobias Krojer,[a] Romain Talon,[a] Nicholas Pearce,[a] Patrick Collins,[b] Alice Douangamath,[b] Jose Brandao-Neto,[b] Alexandre Dias,[b] Brian Marsden[a,c] and Frank von Delft[a,b,d]*

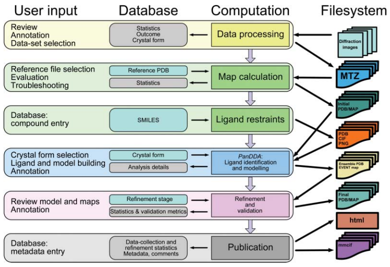
[a]Structural Genomics Consortium, University of Oxford, Roosevelt Drive, Oxford OX3 7DQ, England, [b]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0QX, England, [c]Kennedy Institute of Rheumatology, University of Oxford, Roosevelt Drive, Oxford OX3 7FY, England, and [d]Department of Biochemistry, University of Johannesburg, Auckland Park 2006, South Africa. *Correspondence e-mail: frank.vondelft@sgc.ox.ac.uk

Received 1 November 2016
Accepted 20 December 2016

Keywords: *XChemExplorer*; *PanDDA*; structure-based ligand design; protein–ligand structure; fragment screening.

Supporting information: this article has supporting information at journals.iucr.org/d

XChemExplorer (*XCE*) is a data-management and workflow tool to support large-scale simultaneous analysis of protein–ligand complexes during structure-based ligand discovery (SBLD). The user interfaces of established crystallographic software packages such as *CCP*4 [Winn *et al.* (2011), *Acta Cryst.* D**67**, 235–242] or *PHENIX* [Adams *et al.* (2010), *Acta Cryst.* D**66**, 213–221] have entrenched the paradigm that a 'project' is concerned with solving one structure. This does not hold for SBLD, where many almost identical structures need to be solved and analysed quickly in one batch of work. Functionality to track progress and annotate structures is essential. *XCE* provides an intuitive graphical user interface which guides the user from data processing, initial map calculation, ligand identification and refinement up until data dissemination. It provides multiple entry points depending on the need of each project, enables batch processing of multiple data sets and records metadata, progress and annotations in an SQLite database. *XCE* is freely available and works on any Linux and Mac OS X system, and the only dependency is to have the latest version of *CCP*4 installed. The design and usage of this tool are described here, and its usefulness is demonstrated in the context of fragment-screening campaigns at the Diamond Light Source. It is routinely used to analyse projects comprising 1000 data sets or more, and therefore scales well to even very large ligand-design projects.



1. Introduction

Protein crystallography is a cornerstone of structure-based ligand design (SBLD; Blundell *et al.*, 2002), providing the molecular details of how newly synthesized compounds interact with the protein of interest in order to drive the compound-design process. This requires iterative cycles of design and synthesis of compounds with improved binding characteristics, based on structural information of previous protein–ligand complexes and biophysical measurements (Carvalho *et al.*, 2009). In the past, SBLD was the preserve of industry, but academic groups have lately started to pick up on this approach (Hole *et al.*, 2013; Brem *et al.*, 2016).

Depending on the magnitude of the project and its resources, every development cycle may yield tens of novel compounds, each of which requires a protein–ligand structure, either by co-crystallization or soaking into preformed crystals. The widespread availability of third-generation synchrotron sources which are equipped with reliable robotic sample changers and fast X-ray detectors has made it possible to

https://doi.org/10.1107/S2059798316020234

collect hundreds of data sets during a single visit (Broennimann *et al.*, 2006; Nurizzo *et al.*, 2016). The extreme case is crystallographic fragment screening, in which hundreds of similar data sets are collected rapidly and automatically (Bauman *et al.*, 2013; Schiebel *et al.*, 2016). In 2012, Wasserman and coworkers estimated that industry determines in excess of 10 000 protein–ligand structures per year (Wasserman *et al.*, 2012), and in the light of ever-improving hardware and software this number is likely to increase manyfold in the near future. For instance, the Oxford site of the Structural Genomics Consortium (SGC) alone collected 960 data sets from different protein–ligand complexes in 2015.

The success and throughput of SBLD projects depends on the availability of robust and well diffracting crystal systems (Danley, 2006). Hence, from a crystallographic point of view, data collection, processing and refinement of such samples might appear to be algorithmically trivial: most crystals will be nearly isomorphous, so that phasing is achieved by simple molecular substitution, while adjustments to the model are usually restricted to the region around the respective ligand-binding site.

Crystallographic software packages have undergone a remarkable degree of automation over the last decade which facilitates all steps of structure determination, starting with data processing (Vonrhein *et al.*, 2011; Winter, 2010; Sparta *et al.*, 2016; Grochulski *et al.*, 2012), through phasing (Holton & Alber, 2004; Keegan & Winn, 2008; Ness *et al.*, 2004; Panjikar *et al.*, 2005; Terwilliger *et al.*, 2009) and model building (Cowtan, 2006; Porebski *et al.*, 2016; Terwilliger *et al.*, 2008), all the way to refinement (Joosten *et al.*, 2012). Additionally, pipelines have become available to specifically support the solution of protein–ligand structures (Echols *et al.*, 2014).

However, what remains unaddressed is the logistics of analysing large numbers of data sets. Crystallographic software suites such as *CCP*4 or *PHENIX* have not evolved beyond the early paradigm that one project leads to one structure, a design that is adequate for the determination of novel or otherwise challenging crystal structures but does not really reflect the requirements of SBLD projects. In SBLD, a single project consists of multiple structures, all of which need to be analysed in relation to each other. Additionally, protein–ligand structure solution is usually a linear, well defined process and mostly requires the repetitive usage of a limited number of software tools. This seemingly simple workflow becomes tedious because it requires the user to relaunch the same processes over and over again. Ideally, users should be able to analyse, build and refine multiple structures within one software instance without having to open and close different programs.

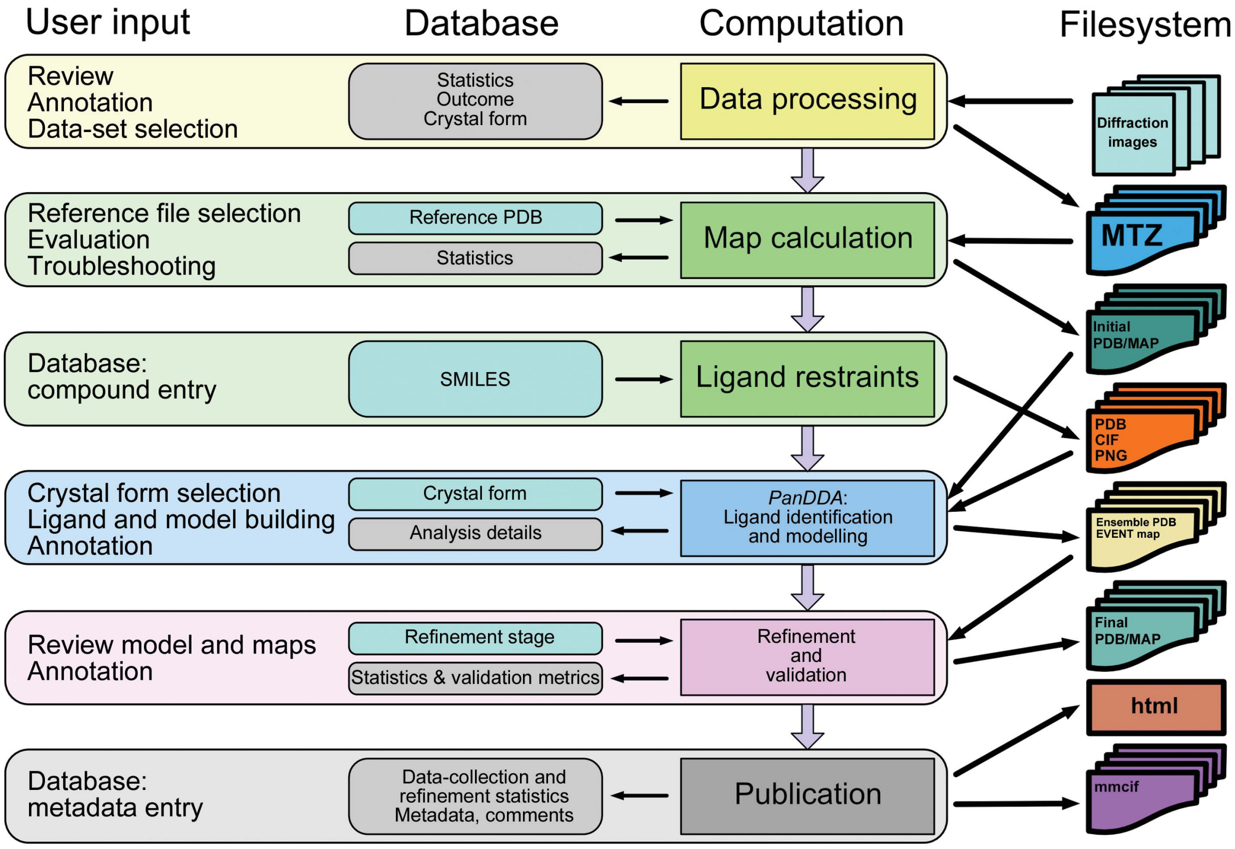


Figure 1
Conceptual design of *XChemExplorer*. The diagram illustrates the supported workflow and the interplay of database and filesystem. User input is mediated through the graphical user interface.

The *AutoSolve* platform from Astex (Mooij *et al.*, 2006) represents one solution to the problem, and it can be assumed that many industrial or industry-like groups have developed similar software solutions that tie in with their internal databases. However, these solutions are not publicly available and, even if they were, such systems would not be trivial to set up or to maintain. There exist no up-to-date workflow-management tools to support multi-data-set analysis that are available to the broad crystallographic community.

Here, we present *XChemExplorer* (*XCE*), a graphical, data-management and workflow tool that brings together established programs that are typically used to process, solve and refine protein–ligand structures. *XCE* does not try to automate the structure-determination process, but it serves as the control centre for running the necessary programs and allows the user to annotate each experiment easily but thoroughly. *XCE* was originally developed to support crystallographic fragment screening at the Diamond Light Source, but it is a generic tool that can be used for any project that needs to solve many similar protein–ligand structures.

2. Philosophy

Protein–ligand structure determination may be an algorithmically simple and linear process, but it turns into a challenge when many samples are analysed as part of an SBLD project which can go on for months and years. Not only are there specific demands on project design, but there have to be mechanisms to capture progress and results. This becomes exponentially challenging as the numbers of data sets become large, unless it is explicitly supported by software. For example, compounds often do not bind to a protein and the reason for this is not always immediately clear. While the experiment ultimately remains unsuccessful, the reason for the failure may be unrelated to the compound, and become evident only when every step of that particular experiment is analysed. *XCE* records compound information, metadata, file locations, observations, outcome and status of the project in a database. It serves as a funnel which channels disparate data into a grid so that it becomes comprehensible for the scientist. The program does not try to make decisions for the crystallographer, but it provides the tools to evaluate each experiment, assign a result and (re)run calculations in batches. The recommended workflow (Fig. 1) is reflected by the design of the graphical user interface (Figs. 2–5): each tab in the interface represent a milestone in the workflow, and the processes necessary to achieve the respective objectives are bundled in the coloured boxes at the bottom of the panel. The main workflow is grouped into four sections (data processing, initial map calculation, hit identification and refinement) and it is possible to enter at any point, provided that the required data are present in a suitable format.

XCE uses a relational SQLite database to store, retrieve and update all of the metadata that are generated as part of a project. The database can be populated by different means depending on the entry point.

(i) In the most general case, *XCE* will be used to process diffraction images, whereby the software will create a sample entry for every data set.

(ii) It is possible to manually create the anticipated directory structure, and the software will be able to auto-populate the database while parsing the project directory. Details of the naming conventions and project directory structure are given in §3.2 and Supporting Information §S1.

(iii) If the data were collected at the Diamond Light Source then the database can be populated by parsing the respective visit directory as long as the data sets were collected according to the following folder structure: `<protein name>/<crystal name>`, *e.g.* `Lysozyme/Lysozyme-x001` *etc*. All results from automatic data processing are stored in a folder named `processed` in the respective visit directory, and the folder hierarchy during data collection is reflected in the structure of the processed folder. It contains a directory for each protein target and subfolders for every protein crystal belonging to the target, which then branch into subfolders for different runs and finally into folders for the different data-processing pipelines.

(iv) In the special case where crystals were mounted and collected as part of a crystallographic fragment-screening project at beamline I04-1 at the Diamond Light Source, then the resulting SQLite file can be directly used as input for *XCE*.

XCE tracks the status of each mounted crystal by storing the data-collection outcome, refinement stage and ligand confidence in the respective field in the database. An overview and an explanation of the different flags can be found in Supplementary Tables S1–S3. Some of the assignments are performed automatically, while others must be explicitly triggered by the user.

3. Software design and implementation

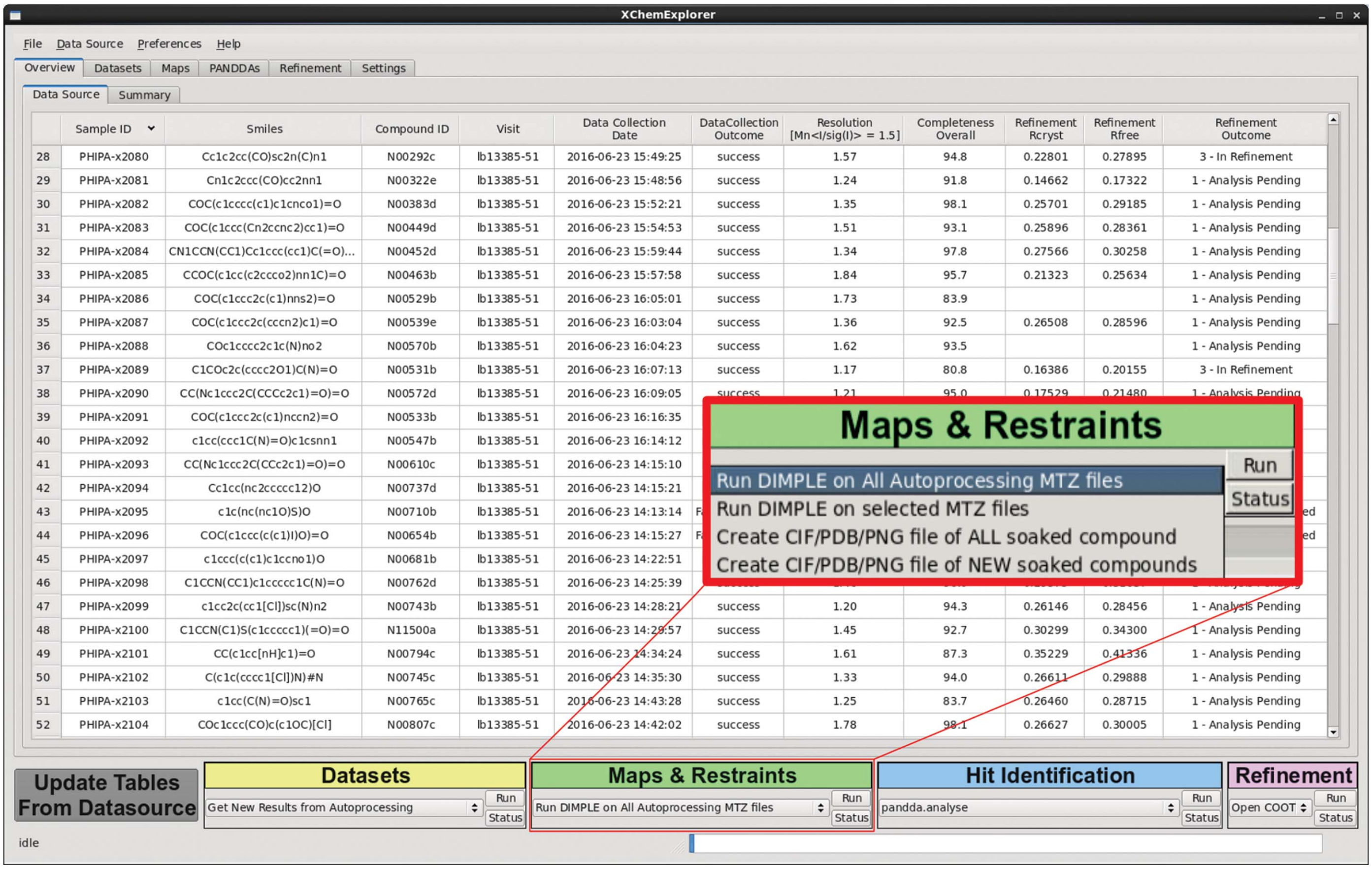
XChemExplorer is implemented in Python (http://www.python.org) and, like *CCP*4, uses the PyQt4 library to provide the graphical user interface functionality. It runs on any Linux or Mac OS X system that has *CCP*4 v.7.0 or higher installed. *XCE* uses SQLite (http://www.sqlite.org) as a relational database-management system. The *RDKit* library (http://www.rdkit.org) is used to create two-dimensional images of compounds. Since *XCE* is a workflow and data-management tool and not an algorithm, it makes use of existing software for different parts of the structure-determination process: *xia*2 (Winter, 2010) for data processing; *DIMPLE* for initial refinement and map calculation; *AceDRG* (Long *et al.*, 2017) for generation of ligand coordinates and restraints; *REFMAC* (Murshudov *et al.*, 2011) for refinement; *PanDDA* (Pearce *et al.*, 2016) for hit identification; a variety of tools from *PHENIX* (Adams *et al.*, 2010) for validation purposes; and *Coot* (Emsley *et al.*, 2010) for model building. There is, however, no fundamental limitation to incorporating other software packages. Source code for *XChemExplorer* and installation instructions are available at http://tkrojer.github.io/XChemExplorer.

3.1. Filesystem organization

XCE is a filesystem-based program, meaning that all files are stored in the same project directory and have to be organized in a hierarchical manner. All files belonging to one crystal are stored in a subfolder and the name of this folder is equal to the sample identifier. The filenames have to adhere to a naming convention, but the program will manage the correct labelling if used throughout the process. Nevertheless, the requirements are minimal and it is easy to add data manually by adjusting the folder and file names to the expected nomenclature. All that needs to be provided are an MTZ file and an *AIMLESS* logfile (Evans & Murshudov, 2013) in the respective sample directory. The files must have the same filename root as the folder name, *e.g.* if the folder name is `TEST-x001`, then the MTZ and logfile must be called `TEST-x001.mtz` and `TEST-x001.log`, respectively. Alternatively, the MTZ file and logfile must have the same name in all subfolders, *e.g.* `datafile.mtz` and `datafile.log`. Details of naming conventions are given in Supporting Information §S1.

3.2. SQLite database

XCE relies on an SQLite relational database to store and retrieve all of the many pieces of metadata for each sample and experiment. As described in §2, the database can be generated in different ways, while its contents can be viewed and edited in the Overview tab (Fig. 2). An SQLite database is essentially just a file and therefore requires little maintenance, in contrast to more sophisticated server–client database engines which are not easy to set up and maintain without dedicated IT support. *XCE* could be configured to support such systems, but priority was given to establishing a transparent, comprehensible and editable system which reflects the day-to-day reality of the typical practising crystallographer. The database consists of a main table which contains a row for each sample and columns for different metadata, and a secondary table to capture results specific to the novel *PanDDA* (*Pan-Dataset Density Analysis*) hit-analysis software (Pearce *et al.*, 2016). The location and name of the database is defined in the settings of *XCE* and, since the database is just a file, it can be located anywhere on the filesystem. Access



	Sample ID	Smiles	Compound ID	Visit	Data Collection Date	DataCollection Outcome	Resolution [Mn<I/sig(I)> = 1.5]	Completeness Overall	Refinement Rcryst	Refinement Rfree	Refinement Outcome
28	PHIPA-x2080	Cc1c2cc(CO)sc2n(C)n1	N00292c	lb13385-51	2016-06-23 15:49:25	success	1.57	94.8	0.22801	0.27895	3 - In Refinement
29	PHIPA-x2081	Cn1c2ccc(CO)cc2nn1	N00322e	lb13385-51	2016-06-23 15:48:56	success	1.24	91.8	0.14662	0.17322	1 - Analysis Pending
30	PHIPA-x2082	COC(c1cccc(c1)c1cnco1)=O	N00383d	lb13385-51	2016-06-23 15:52:21	success	1.35	98.1	0.25701	0.29185	1 - Analysis Pending
31	PHIPA-x2083	COC(c1ccc(Cn2ccnc2)cc1)=O	N00449d	lb13385-51	2016-06-23 15:54:53	success	1.51	93.1	0.25896	0.28361	1 - Analysis Pending
32	PHIPA-x2084	CN1CCN(CC1)Cc1ccc(cc1)C(=O)...	N00452d	lb13385-51	2016-06-23 15:59:44	success	1.34	97.8	0.27566	0.30258	1 - Analysis Pending
33	PHIPA-x2085	CCOC(c1cc(c2ccco2)nn1C)=O	N00463b	lb13385-51	2016-06-23 15:57:58	success	1.84	95.7	0.21323	0.25634	1 - Analysis Pending
34	PHIPA-x2086	COC(c1ccc2c(c1)nns2)=O	N00529b	lb13385-51	2016-06-23 16:05:01	success	1.73	83.9			1 - Analysis Pending
35	PHIPA-x2087	COC(c1ccc2c(cccn2)c1)=O	N00539e	lb13385-51	2016-06-23 16:03:04	success	1.36	92.5	0.26508	0.28596	1 - Analysis Pending
36	PHIPA-x2088	COc1cccc2c1c(N)no2	N00570b	lb13385-51	2016-06-23 16:04:23	success	1.62	93.5			1 - Analysis Pending
37	PHIPA-x2089	C1COc2c(cccc2O1)C(N)=O	N00531b	lb13385-51	2016-06-23 16:07:13	success	1.17	80.8	0.16386	0.20155	3 - In Refinement
38	PHIPA-x2090	CC(Nc1ccc2C(CCCc2c1)=O)=O	N00572d	lb13385-51	2016-06-23 16:09:05	success	1.21	95.0	0.17529	0.21480	1 - Analysis Pending
39	PHIPA-x2091	COC(c1ccc2c(c1)nccn2)=O	N00533b	lb13385-51	2016-06-23 16:16:35						
40	PHIPA-x2092	c1cc(ccc1C(N)=O)c1csnn1	N00547b	lb13385-51	2016-06-23 16:14:12						
41	PHIPA-x2093	CC(Nc1ccc2C(CCc2c1)=O)=O	N00610c	lb13385-51	2016-06-23 14:15:10						
42	PHIPA-x2094	Cc1cc(nc2ccccc12)O	N00737d	lb13385-51	2016-06-23 14:15:21						
43	PHIPA-x2095	c1c(nc(nc1O)S)O	N00710b	lb13385-51	2016-06-23 14:13:14						
44	PHIPA-x2096	COC(c1ccc(c(c1)I)O)=O	N00654b	lb13385-51	2016-06-23 14:15:27						
45	PHIPA-x2097	c1ccc(c(c1)c1ccno1)O	N00681b	lb13385-51	2016-06-23 14:22:51						
46	PHIPA-x2098	C1CCN(CC1)c1ccccc1C(N)=O	N00762d	lb13385-51	2016-06-23 14:25:39						
47	PHIPA-x2099	c1cc2c(cc1[Cl])sc(N)n2	N00743b	lb13385-51	2016-06-23 14:28:21	success	1.20	94.3	0.26146	0.28456	1 - Analysis Pending
48	PHIPA-x2100	C1CCN(C1)S(c1ccccc1)(=O)=O	N11500a	lb13385-51	2016-06-23 14:29:57	success	1.45	92.7	0.30299	0.34300	1 - Analysis Pending
49	PHIPA-x2101	CC(c1cc[nH]c1)=O	N00794c	lb13385-51	2016-06-23 14:34:24	success	1.61	87.3	0.35229	0.41336	1 - Analysis Pending
50	PHIPA-x2102	C(c1c(cccc1[Cl])N)#N	N00745c	lb13385-51	2016-06-23 14:35:30	success	1.33	94.0	0.26611	0.29888	1 - Analysis Pending
51	PHIPA-x2103	c1cc(C(N)=O)sc1	N00765c	lb13385-51	2016-06-23 14:43:28	success	1.25	83.7	0.26460	0.28715	1 - Analysis Pending
52	PHIPA-x2104	COc1ccc(CO)c(c1OC)[Cl]	N00807c	lb13385-51	2016-06-23 14:42:02	success	1.78	98.1	0.26627	0.30005	1 - Analysis Pending

Maps & Restraints
Run DIMPLE on All Autoprocessing MTZ files
Run DIMPLE on selected MTZ files
Create CIF/PDB/PNG file of ALL soaked compound
Create CIF/PDB/PNG file of NEW soaked compounds
Run
Status
Update Tables From Datasource
Datasets
Get New Results from Autoprocessing
Run
Status
Maps & Restraints
Run DIMPLE on All Autoprocessing MTZ files
Run
Status
Hit Identification
pandda.analyse
Run
Status
Refinement
Open COOT
Run
Status
idle

Figure 2
Overview and design principle of the *XChemExplorer* GUI. The Overview tab shows the contents of the database. The general design principle is the same for all tabs: information and parameters which are specific for each milestone during structure solution are displayed in the centre. The coloured action boxes at the bottom of the window bundle a set of commands that are relevant for each stage of the structure-determination process. The inset gives an example of the different processes that can be executed in the Maps & Restraints tab. The grey button at the bottom left corner is for refreshing the tables of the GUI. Status messages and a progress bar in the lowermost part of the window provide feedback and inform about ongoing processes.

control is file-based, so that any user with read-access to the relevant file system will be able to access the data. The content of the database can be displayed with *XCE* or with freely available software such as the *SQLite Browser* (http://sqlitebrowser.org)

XCE can export the database to a comma-separated file (csv), which can then be read into programs such as *Microsoft Excel* for analysis or manipulation. An updated csv file can be reimported into the database if necessary, which is particularly useful for retroactively adding ligand information. *XCE* does not work without a database being specified, but it is always possible to recreate the database from an existing filesystem. In its current form, *XCE* does not provide an interface to external databases, but users who have their own database can easily query the SQLite file and synchronize it with their own system.

3.3. Job submission

XCE checks for the availability of a functional queuing system during startup and will submit all computationally intensive calculations to the cluster. Alternatively, the local machine will be used. At the moment only the Portable Batch System (PBS) is supported, but implementation of other queuing systems would not in principle be difficult.

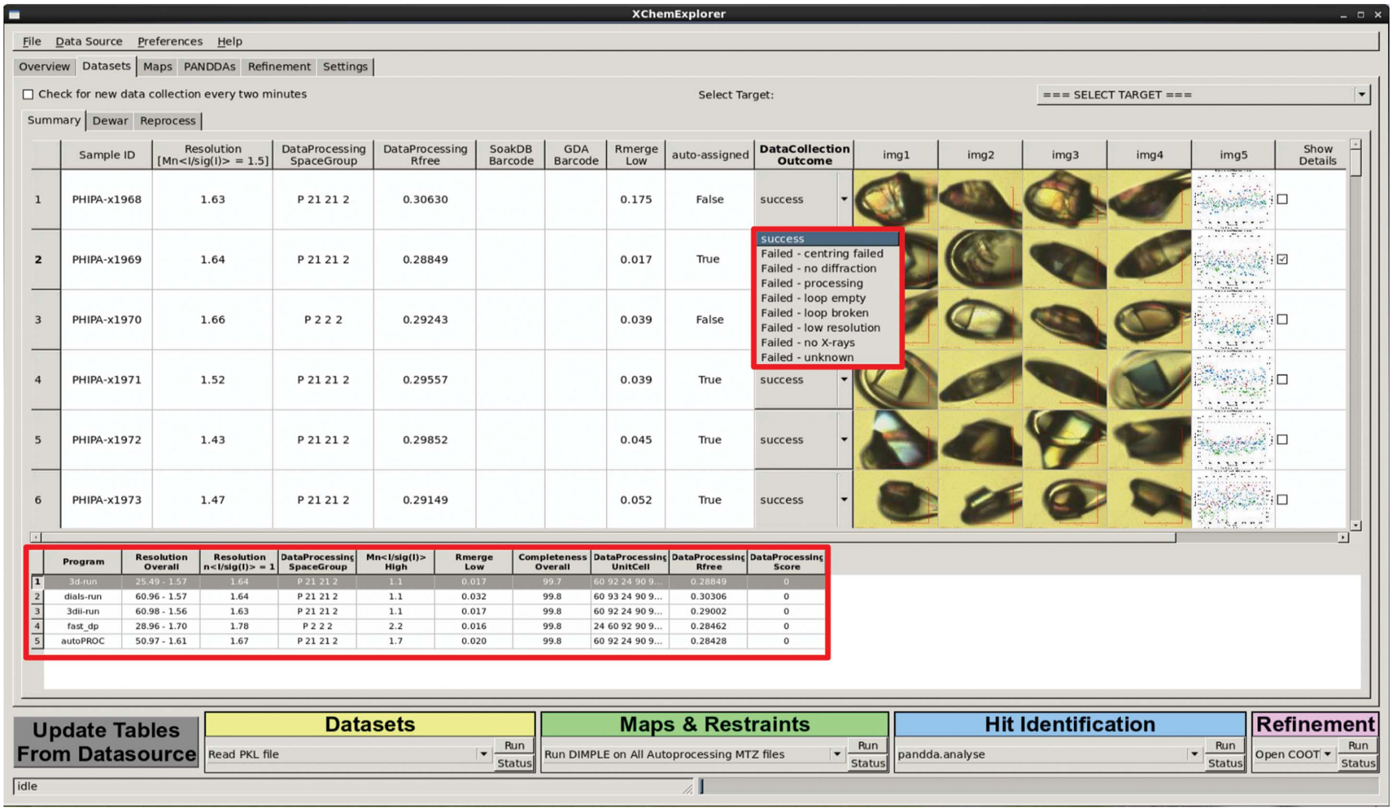
4. Description of usage

4.1. Starting the program

The program can be started with little preparation. It is sufficient to provide a database file and the location of a project directory. If the database does not already exist, then the 'Data Source' menu can be used to create a database from scratch. Additional directories can be specified as relevant, for example the location of reference files, the output of *PanDDA* hit analysis or a visit directory at the Diamond Light Source.

4.2. Data-collection review and data (re)processing

The Datasets tab is for reviewing data collections at the Diamond Light Source and for analysing data sets processed with different auto-processing pipelines. Additionally, diffraction images can be (re)processed with the *xia*2 data-reduction system (Winter, 2010). Historically, this was the first module of the program, because data collected at Diamond Light Source are immediately processed by several automated data-processing pipelines (Winter & McAuley, 2011). The availability of such a system appears to be indispensable for any ambitious SBLD program since the multiple programs provide redundancy that ensures that the pipeline robustly provides outputs. Individual programs are prone to



□ Check for new data collection every two minutes

Select Target: === SELECT TARGET ===

	Sample ID	Resolution [Mn<I/sig(I)> = 1.5]	DataProcessing SpaceGroup	DataProcessing Rfree	SoakDB Barcode	GDA Barcode	Rmerge Low	auto-assigned	DataCollection Outcome	Show Details
1	PHIPA-x1968	1.63	P 21 21 2	0.30630			0.175	False	success	□
2	PHIPA-x1969	1.64	P 21 21 2	0.28849			0.017	True		☑
3	PHIPA-x1970	1.66	P 2 2 2	0.29243			0.039	False		□
4	PHIPA-x1971	1.52	P 21 21 2	0.29557			0.039	True	success	□
5	PHIPA-x1972	1.43	P 21 21 2	0.29852			0.045	True	success	□
6	PHIPA-x1973	1.47	P 21 21 2	0.29149			0.052	True	success	□

success
Failed - centring failed
Failed - no diffraction
Failed - processing
Failed - loop empty
Failed - loop broken
Failed - low resolution
Failed - no X-rays
Failed - unknown

	Program	Resolution Overall	Resolution n<I/sig(I)> = 1	DataProcessing SpaceGroup	Mn<I/sig(I)> High	Rmerge Low	Completeness Overall	DataProcessing UnitCell	DataProcessing Rfree	DataProcessing Score
1	3d-run	25.49 - 1.57	1.64	P 21 21 2	1.1	0.017	99.7	60 92 24 90 9...	0.28849	0
2	dials-run	60.96 - 1.57	1.64	P 21 21 2	1.1	0.032	99.8	60 93 24 90 9...	0.30306	0
3	3dii-run	60.98 - 1.56	1.63	P 21 21 2	1.1	0.017	99.8	60 92 24 90 9...	0.29002	0
4	fast_dp	28.96 - 1.70	1.78	P 2 2 2	2.2	0.016	99.8	24 60 92 90 9...	0.28462	0
5	autoPROC	50.97 - 1.61	1.67	P 21 21 2	1.7	0.020	99.8	60 92 24 90 9...	0.28428	0

Datasets: Read PKL file

Maps & Restraints: Run DIMPLE on All Autoprocessing MTZ files

Hit Identification: pandda.analyse

Refinement: Open COOT

idle

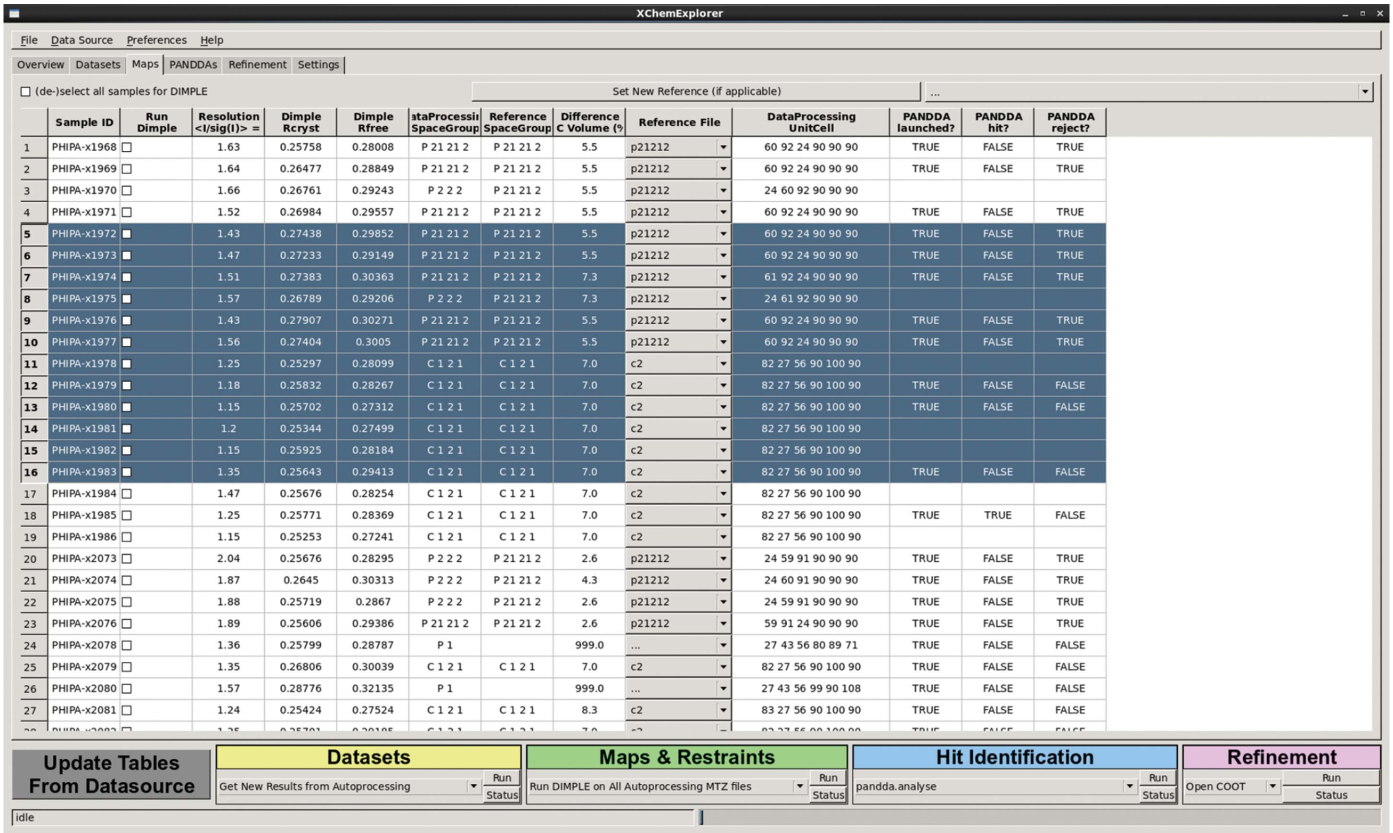
Figure 3
The Datasets panel which allows users to review data collected at the Diamond Light Source. Selected data-collection statistics, crystal images after alignment and a drop-down widget for annotating the outcome of the experiment are displayed for each crystal. The upper red box shows a selection of the various data-collection outcome annotations. The Show Details checkbox is for displaying an extended view of the output from all auto-processing pipelines; the lower red box shows detailed data-collection statistics for different auto-processing pipelines. It is possible to manually choose any outcome from the respective table.

unpredictable failures in ways that are neither feasible nor interesting to troubleshoot in the face of large amounts of raw data. At the same time, it does present the challenge of having to assess and select one of multiple processed versions of the data for each crystal.

After the processed data are available, the program will walk through the data directories, check for the presence of scaled MTZ files and parse the corresponding *AIMLESS* logfiles. For data collected at Diamond Light Source, it will additionally read and encode crystal images which are recorded before data collection and which indicate whether the specimen was correctly aligned. MTZ and logfiles from all auto-processing pipelines are then immediately copied to the project directory. The program can additionally monitor an ongoing data collection by parsing a specified visit directory every 2 min and dynamically updating the table in the Datasets tab. The current implementation of *XCE* allows only data collected at Diamond Light Source to be reviewed, because the specific directory structure at Diamond is necessarily hard-coded. However, the relevant code is isolated, meaning it can be adapted for other sites with relative ease.

Results are displayed in the central table, showing one line for each sample, together with the currently selected processing results, crystal images after alignment and a drop-down list that specifies the outcome of the experiment (Fig. 3). *XCE* will copy the final MTZ and logfiles from all pipelines into the project directory, but only the one that is currently selected will be used for any downstream calculations. The program will preselect one result, but it is possible to expand the view and to manually select any data-processing result. Details of the available selection criteria can be found in Supporting Information §S2.

The annotation mechanism is implemented by a drop-down menu for each sample which can be set with a quick scroll of the mouse wheel: this operational simplicity is a vital prerequisite for data sets to be annotated completely. *XCE* automatically flags data collections as successful if a valid MTZ file is found which fulfils a user-defined resolution criterion. If no MTZ file can be found in a given sample directory then *XCE* will label the data collection as an unknown failure. However, if a user wants to specify the type of failure in more detail, for example no diffraction, loop broken *etc.*, then they can do so by manually changing the text of the drop-down menu (Fig. 3). Details about the available annotation criteria are given in Supplementary Table S1. The snapshots after crystal alignment are essential for determining



□ (de-)select all samples for DIMPLE

Set New Reference (if applicable) ...

	Sample ID	Run Dimple	Resolution <I/sig(I)> =	Dimple Rcryst	Dimple Rfree	ataProcessii SpaceGroup	Reference SpaceGroup	Difference C Volume (%	Reference File	DataProcessing UnitCell	PANDDA launched?	PANDDA hit?	PANDDA reject?
1	PHIPA-x1968	□	1.63	0.25758	0.28008	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
2	PHIPA-x1969	□	1.64	0.26477	0.28849	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
3	PHIPA-x1970	□	1.66	0.26761	0.29243	P 2 2 2	P 21 21 2	5.5	p21212	24 60 92 90 90 90			
4	PHIPA-x1971	□	1.52	0.26984	0.29557	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
5	PHIPA-x1972	□	1.43	0.27438	0.29852	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
6	PHIPA-x1973	□	1.47	0.27233	0.29149	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
7	PHIPA-x1974	□	1.51	0.27383	0.30363	P 21 21 2	P 21 21 2	7.3	p21212	61 92 24 90 90 90	TRUE	FALSE	TRUE
8	PHIPA-x1975	□	1.57	0.26789	0.29206	P 2 2 2	P 21 21 2	7.3	p21212	24 61 92 90 90 90			
9	PHIPA-x1976	□	1.43	0.27907	0.30271	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
10	PHIPA-x1977	□	1.56	0.27404	0.3005	P 21 21 2	P 21 21 2	5.5	p21212	60 92 24 90 90 90	TRUE	FALSE	TRUE
11	PHIPA-x1978	□	1.25	0.25297	0.28099	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90			
12	PHIPA-x1979	□	1.18	0.25832	0.28267	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90	TRUE	FALSE	FALSE
13	PHIPA-x1980	□	1.15	0.25702	0.27312	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90	TRUE	FALSE	FALSE
14	PHIPA-x1981	□	1.2	0.25344	0.27499	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90			
15	PHIPA-x1982	□	1.15	0.25925	0.28184	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90			
16	PHIPA-x1983	□	1.35	0.25643	0.29413	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90	TRUE	FALSE	FALSE
17	PHIPA-x1984	□	1.47	0.25676	0.28254	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90			
18	PHIPA-x1985	□	1.25	0.25771	0.28369	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90	TRUE	TRUE	FALSE
19	PHIPA-x1986	□	1.15	0.25253	0.27241	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90			
20	PHIPA-x2073	□	2.04	0.25676	0.28295	P 2 2 2	P 21 21 2	2.6	p21212	24 59 91 90 90 90	TRUE	FALSE	TRUE
21	PHIPA-x2074	□	1.87	0.2645	0.30313	P 2 2 2	P 21 21 2	4.3	p21212	24 60 91 90 90 90	TRUE	FALSE	TRUE
22	PHIPA-x2075	□	1.88	0.25719	0.2867	P 2 2 2	P 21 21 2	2.6	p21212	24 59 91 90 90 90	TRUE	FALSE	TRUE
23	PHIPA-x2076	□	1.89	0.25606	0.29386	P 21 21 2	P 21 21 2	2.6	p21212	59 91 24 90 90 90	TRUE	FALSE	TRUE
24	PHIPA-x2078	□	1.36	0.25799	0.28787	P 1		999.0	...	27 43 56 80 89 71	TRUE	FALSE	FALSE
25	PHIPA-x2079	□	1.35	0.26806	0.30039	C 1 2 1	C 1 2 1	7.0	c2	82 27 56 90 100 90	TRUE	FALSE	FALSE
26	PHIPA-x2080	□	1.57	0.28776	0.32135	P 1		999.0	...	27 43 56 99 90 108	TRUE	FALSE	FALSE
27	PHIPA-x2081	□	1.24	0.25424	0.27524	C 1 2 1	C 1 2 1	8.3	c2	83 27 56 90 100 90	TRUE	FALSE	FALSE

Figure 4
The Maps & Restraints panel which allows users to calculate initial maps and generate ligand coordinates and restraints. The central table gives an overview of all data sets that can be used for initial map calculation with *DIMPLE*. $R_{\rm cryst}$ and $R_{\rm free}$ are displayed if initial refinement has already taken place. Like all other tables, the table can be sorted by any of the displayed columns. *XCE* tries to automatically assign a suitable reference file for input into *DIMPLE*. The drop-down menu displays the currently selected reference file, but it is possible to select any other file from the list. The last three columns indicate the status of each data set, *i.e.* whether they were already subject to analysis and inspection with the *PanDDA* algorithm. Ranges of rows can be selected and marked for map calculation with *DIMPLE*.

the reason for failure: correct crystal alignment is only a peripheral problem for manual data collection, but can be a major issue for automatically centred crystals.

The final facilitator of identifying problematic crystals is the ability to sort the table by any of the displayed categories. Thus, for instance, abnormally large $R_{\rm merge}$ values in the inner resolution shell of data sets are usually a good indicator of samples that warrant closer inspection.

4.3. Initial map calculation

The Maps & Restraints panel enables initial refinement of each data set with the *DIMPLE* difference-map pipeline (Fig. 4). The table provides an overview of all successfully reduced data sets, their resolution and $R_{\rm cryst}/R_{\rm free}$ values in the case where any refinement has been carried out already. *XCE* uses a simple but fast and effective mechanism to search for suitable PDB input files: it will compare the point group and unit-cell volume of each PDB file in the reference directory with the sample MTZ file and suggest it as input for *DIMPLE* if they have the same point group and differ by less than 10% in unit-cell volume. The latter parameter can be adjusted in the Preferences menu. The currently selected reference file is displayed in the drop-down menu, but any of the files in the Reference directory can be chosen. Individual samples or batches can be selected for (re)processing with *DIMPLE*. Since *DIMPLE* is able to solve the structures of non-isomorphous crystals by performing an additional molecular-replacement step, it may not seem necessary to devote much thought to the careful selection of a suitable reference file. However, it is not advisable to completely rely on automated decision making because it is invariably a warning sign if no suitable reference file is available: the program after all assumes that all data sets are similar to the available reference models and it is always better to prepare a good model outside the *XCE* workflow which accurately explains the respective crystal rather than to continue with a suboptimal model. The latter option may sound tempting at the time, but usually results in additional work downstream.

4.4. Ligand preparation

It is a hallmark of SBLD that novel compounds are used that are unlikely to feature in the Cambridge Structural Database or the Protein Data Bank. Hence, ligand coordinates and accurate geometry restraints need to be generated for subsequent model building and refinement purposes (Kleywegt, 2007). The Maps & Restraints tab allows the calculation of ligand coordinates, restraints and two-dimensional pictures of batches of compounds with *AceDRG* and the *RDkit* library.

Information about the compounds needs to be provided in the SQLite database as a SMILES notation. The resulting files are stored in the respective sample folder in the project directory. The database is updated with information about the location of ligand coordinates and restraints once the files have been successfully created. In case of problems, *XCE* does not analyse the cause for failure, but users can query the database and obtain a quick overview of all failed attempts to generate restraints and find the location of the respective sample folder. Note that ligand coordinates and restraints are automatically loaded into *Coot* during the Refinement stage. Hence, validation and editing functionalities in *Coot* can be used in case the respective values need to be updated (Debreczeni & Emsley, 2012). Not currently supported are scenarios where the chirality of the compound is unknown, which are not uncommon in ligand-binding studies. In such cases the corresponding files need to be generated manually and copied into the respective folder.

4.5. Hit identification (*PanDDA*)

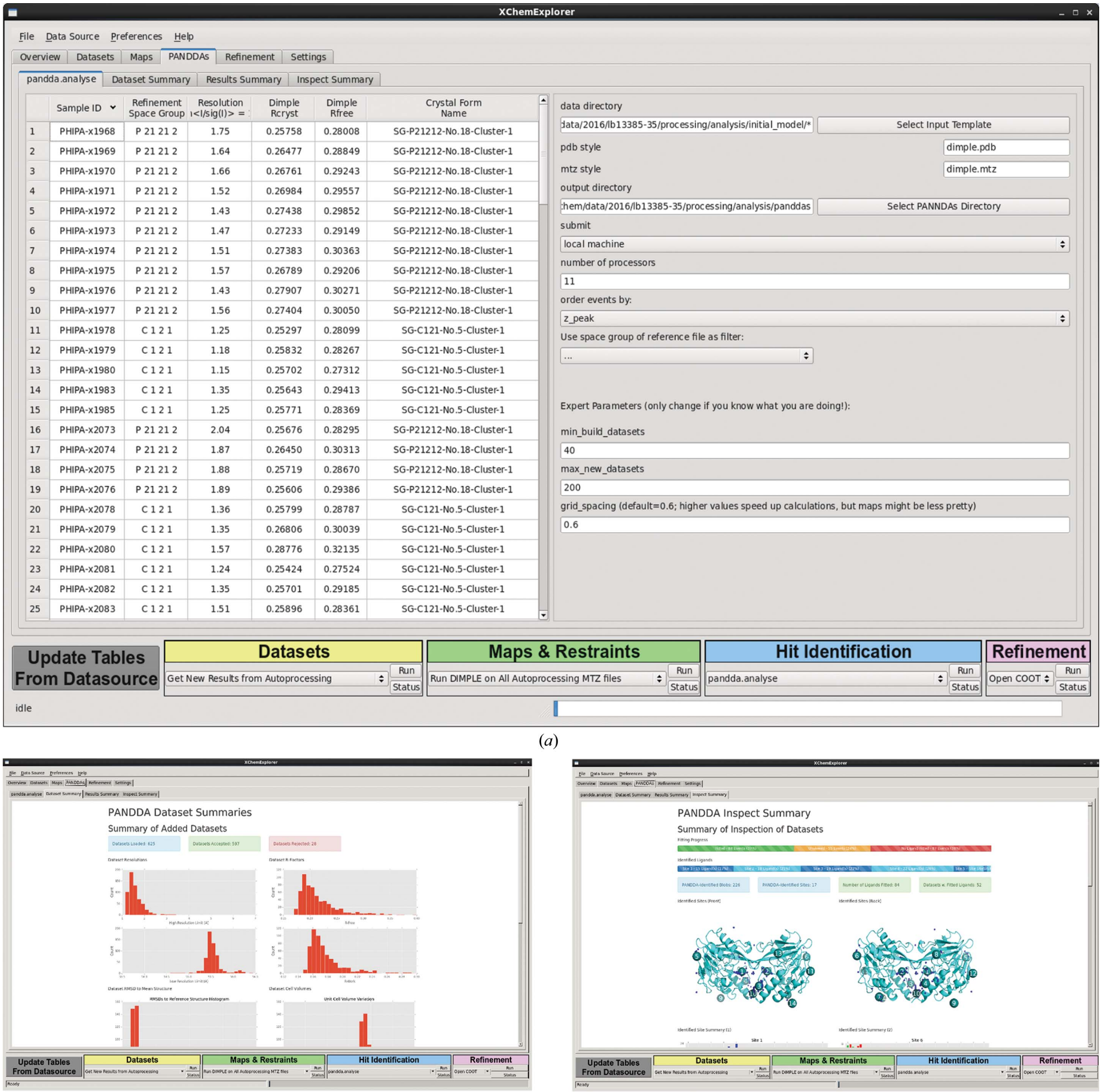
XCE offers a graphical entry point for the recently developed *PanDDA* (*Pan-Dataset Density Analysis*) software (Pearce *et al.*, 2016) for improved hit identification (Fig. 5). Interpreting weak and ambiguous ligand density can significantly slow down a project and make the final models less reliable and dependent on subjective decisions of the responsible scientist (Pozharski *et al.*, 2013), whereas the *PanDDA* algorithm establishes statistically significant measures of confidence for each analysed sample.

The *PanDDA* protocol consists of three steps: analysis, inspection and export. *XCE* displays details of the available data sets for *pandda.analyse* in a table and allows the user to adjust key parameters of the analysis protocol. The results can be displayed with *pandda.inspect*, which uses *Coot* and a special inspection panel for assessing and modelling of the bound ligands. The resulting models undergo a round of refinement and validation during the export step. Additionally, all output which is generated by the *PanDDA* software will be imported into the database at this point. Details of the processes taking place during the export step are given in Supplementary Fig. S2. While it is recommended to use *PanDDA* for hit identification, it is not necessary to be able to continue with the workflow since it is possible to proceed directly to refinement after initial map calculation.

4.6. Refinement

Completing structures requires iterative cycles of model building, refinement and validation, with progress monitored by the convergence of global quality metrics. Unfortunately, the endpoint is often ill-defined and mostly depends on the assessment of a skilled crystallographer. *XCE* addresses this challenge by minimizing the number of cycles through assisting the user in obtaining the best possible starting model (see §§4.3 and 4.5).

The process is greatly accelerated by a graphical command interface which allows modelling and refinement of all data sets within a single instance of *Coot* (Fig. 6). A subset of all available structures can be selected based on their refinement stage and, in the case where the ligands were built with *PanDDA*, location of the respective binding site. The interface displays global and ligand-specific quality metrics of the currently loaded structure, as well as a two-dimensional picture of the ligand and a radar plot summarizing a set of ligand-validation scores: correlation between model and observed electron density (RSCC), a statistical measure of difference density in this region of the model (RSZD), the *B*-factor ratio of ligand and surrounding protein side-chain atoms within 4 Å (*B*-factor ratio), the root-mean-squared deviation of ligand coordinates in the initial model and after refinement (r.m.s.d.), and the strength of density over model normalized for occupancy (RSZO/OCC) (Pearce & von Delft, 2017; Tickle, 2012). The corresponding ligand coordinates and restraints are loaded as a separate molecule in *Coot* when a new molecule is chosen. In the case where the ligand had not been modelled before, then the interface can be used to first place the molecule into the centre of the screen for further fitting and then to merge it into the loaded protein structure. Refinement can be triggered with a single button click and is carried out with *REFMAC*, followed by comprehensive



	Sample ID	Refinement Space Group	Resolution ı<I/sig(I)> =	Dimple Rcryst	Dimple Rfree	Crystal Form Name
1	PHIPA-x1968	P 21 21 2	1.75	0.25758	0.28008	SG-P21212-No.18-Cluster-1
2	PHIPA-x1969	P 21 21 2	1.64	0.26477	0.28849	SG-P21212-No.18-Cluster-1
3	PHIPA-x1970	P 21 21 2	1.66	0.26761	0.29243	SG-P21212-No.18-Cluster-1
4	PHIPA-x1971	P 21 21 2	1.52	0.26984	0.29557	SG-P21212-No.18-Cluster-1
5	PHIPA-x1972	P 21 21 2	1.43	0.27438	0.29852	SG-P21212-No.18-Cluster-1
6	PHIPA-x1973	P 21 21 2	1.47	0.27233	0.29149	SG-P21212-No.18-Cluster-1
7	PHIPA-x1974	P 21 21 2	1.51	0.27383	0.30363	SG-P21212-No.18-Cluster-1
8	PHIPA-x1975	P 21 21 2	1.57	0.26789	0.29206	SG-P21212-No.18-Cluster-1
9	PHIPA-x1976	P 21 21 2	1.43	0.27907	0.30271	SG-P21212-No.18-Cluster-1
10	PHIPA-x1977	P 21 21 2	1.56	0.27404	0.30050	SG-P21212-No.18-Cluster-1
11	PHIPA-x1978	C 1 2 1	1.25	0.25297	0.28099	SG-C121-No.5-Cluster-1
12	PHIPA-x1979	C 1 2 1	1.18	0.25832	0.28267	SG-C121-No.5-Cluster-1
13	PHIPA-x1980	C 1 2 1	1.15	0.25702	0.27312	SG-C121-No.5-Cluster-1
14	PHIPA-x1983	C 1 2 1	1.35	0.25643	0.29413	SG-C121-No.5-Cluster-1
15	PHIPA-x1985	C 1 2 1	1.25	0.25771	0.28369	SG-C121-No.5-Cluster-1
16	PHIPA-x2073	P 21 21 2	2.04	0.25676	0.28295	SG-P21212-No.18-Cluster-1
17	PHIPA-x2074	P 21 21 2	1.87	0.26450	0.30313	SG-P21212-No.18-Cluster-1
18	PHIPA-x2075	P 21 21 2	1.88	0.25719	0.28670	SG-P21212-No.18-Cluster-1
19	PHIPA-x2076	P 21 21 2	1.89	0.25606	0.29386	SG-P21212-No.18-Cluster-1
20	PHIPA-x2078	C 1 2 1	1.36	0.25799	0.28787	SG-C121-No.5-Cluster-1
21	PHIPA-x2079	C 1 2 1	1.35	0.26806	0.30039	SG-C121-No.5-Cluster-1
22	PHIPA-x2080	C 1 2 1	1.57	0.28776	0.32135	SG-C121-No.5-Cluster-1
23	PHIPA-x2081	C 1 2 1	1.24	0.25424	0.27524	SG-C121-No.5-Cluster-1
24	PHIPA-x2082	C 1 2 1	1.35	0.25701	0.29185	SG-C121-No.5-Cluster-1
25	PHIPA-x2083	C 1 2 1	1.51	0.25896	0.28361	SG-C121-No.5-Cluster-1

data directory

data/2016/lb13385-35/processing/analysis/initial_model/* [Select Input Template]

pdb style: dimple.pdb

mtz style: dimple.mtz

output directory

:hem/data/2016/lb13385-35/processing/analysis/panddas [Select PANNDAs Directory]

submit

local machine

number of processors

11

order events by:

z_peak

Use space group of reference file as filter:

...

Expert Parameters (only change if you know what you are doing!):

min_build_datasets

40

max_new_datasets

200

grid_spacing (default=0.6; higher values speed up calculations, but maps might be less pretty)

0.6

(*a*)

PANDDA Dataset Summaries

Summary of Added Datasets

(*b*)

PANDDA Inspect Summary

Summary of Inspection of Datasets

(*c*)

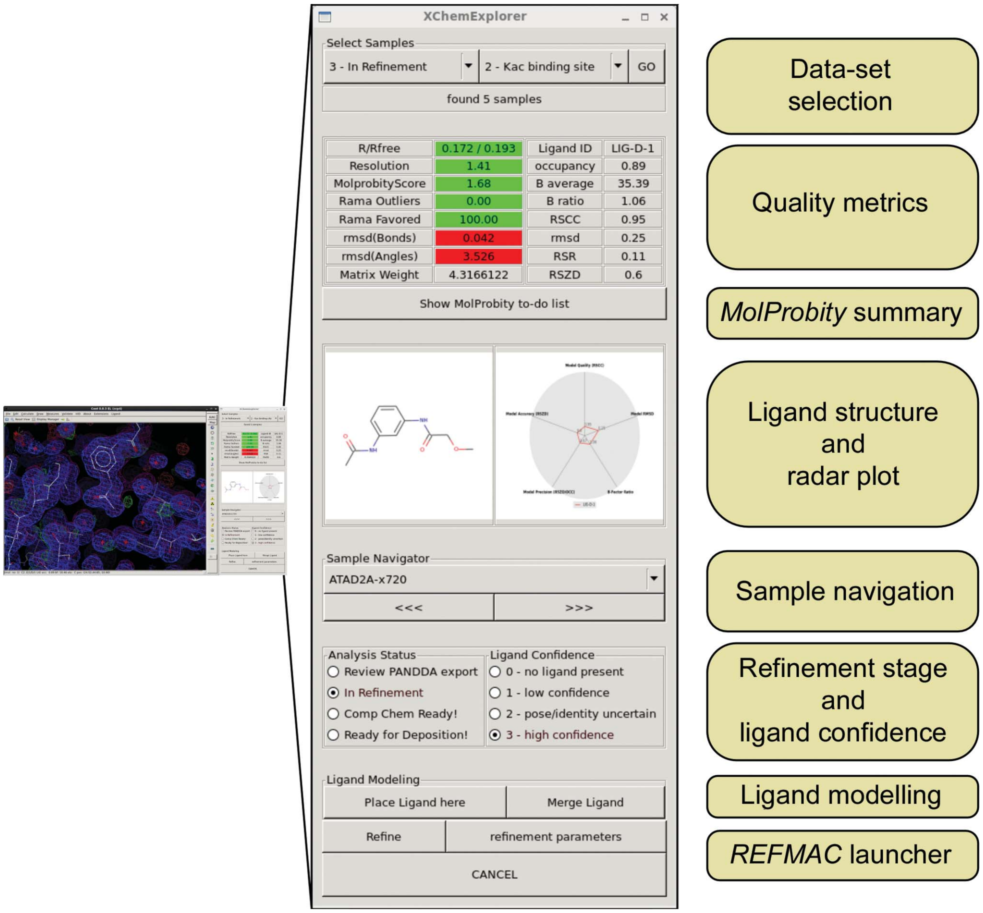
Figure 5
The PanDDA panel allows the configuration of calculations by the *PanDDA* software. (*a*) The left table provides an overview of all the data sets that are available for input into *PanDDA*; the right side displays the currently selected input parameters for *pandda.analyse*. These are relevant mostly when multiple crystal forms are available. (*b*) *PanDDA* data-set summary. (*c*) *PanDDA* inspection summary.

validation with different programs from the *PHENIX* package. Details of the procedure can be found in Supplementary Fig. S3. Finally, global and ligand-specific quality metrics of the resulting model are determined and the database is updated accordingly.

Additionally, a qualitative assessment of refinement stage and ligand confidence of each sample can be chosen from predefined categories. *XCE* uses a hierarchical annotation system to help the refiner describe the refinement stage of a given data set, facilitating the organization and prioritization of samples within a project. The respective database field is updated automatically during data collection, initial refinement with *DIMPLE* and *PanDDA* analysis, but once the final refinement stage is reached crystallographers need to change the flag individually. The categories are mostly self-descriptive and details can be found in Supplementary Table S2. Introduction of this hierarchy helps to provide an overview of the work that is still needed to finish all the models, and it allows subsets of all available structures to be selected which need more work to be performed before they are ready for subsequent analysis and deposition. The ligand confidence category reflects the trust of the refiner in the modelled ligand. This may change as refinement progresses and is informed by how well the electron density or the *PanDDA* event map describes the ligand and by the quantitative metrics described above. Ultimately, it remains a subjective statement, but it is especially useful in cases where the crystallographer cannot unambiguously establish the binding pose and therefore needs a way to communicate the limitations of the model to whoever uses it for further analysis.

4.7. Data sharing

The final problem addressed by *XCE* is how to communicate the results with other scientists; this is of particular importance in SBLD projects, which are per definition multidisciplinary. The software tools and quality metrics that are well known to the crystallographic community tend to be alien to non-structural biologists, and not in fact relevant to their problems. In our experience, they ask the following questions.



XChemExplorer

Select Samples: 3 - In Refinement ▾ | 2 - Kac binding site ▾ | GO

found 5 samples

R/Rfree	🟩 0.172 / 0.193	Ligand ID	LIG-D-1
Resolution	🟩 1.41	occupancy	0.89
MolprobityScore	🟩 1.68	B average	35.39
Rama Outliers	🟩 0.00	B ratio	1.06
Rama Favored	🟩 100.00	RSCC	0.95
rmsd(Bonds)	🟥 0.042	rmsd	0.25
rmsd(Angles)	🟥 3.526	RSR	0.11
Matrix Weight	4.3166122	RSZD	0.6

Show MolProbity to-do list

Sample Navigator: ATAD2A-x720 ▾

<<< | >>>

Analysis Status
- ○ Review PANDDA export
- ● In Refinement
- ○ Comp Chem Ready!
- ○ Ready for Deposition!

Ligand Confidence
- ○ 0 - no ligand present
- ○ 1 - low confidence
- ○ 2 - pose/identity uncertain
- ● 3 - high confidence

Ligand Modeling: Place Ligand here | Merge Ligand

Refine | refinement parameters

CANCEL

Figure 6
The *XCE* Refinement interface for accelerated model building and refinement. The small panel on the left side shows how *Coot* and the new dialogue work in relation to each other; the middle and right side show an extended image of the dialogue and a schematic outline of its arrangement. The dialogue displays global and ligand-specific quality metrics of the currently loaded structure as well as a two-dimensional picture of the ligand and a radar plot summarizing a set of ligand-validation scores. *XCE* will automatically display $2mF_o - DF_c$ and $mF_o - DF_c$ maps as well as *PanDDA* event maps when they are available. The arrow buttons and a drop-down menu in the control panel are used to step through the data sets. It also allows the refinement stage of a given structure and the confidence of the modelled ligand to be assigned. All structures which are part of the project can be handled within a single instance of *Coot*.

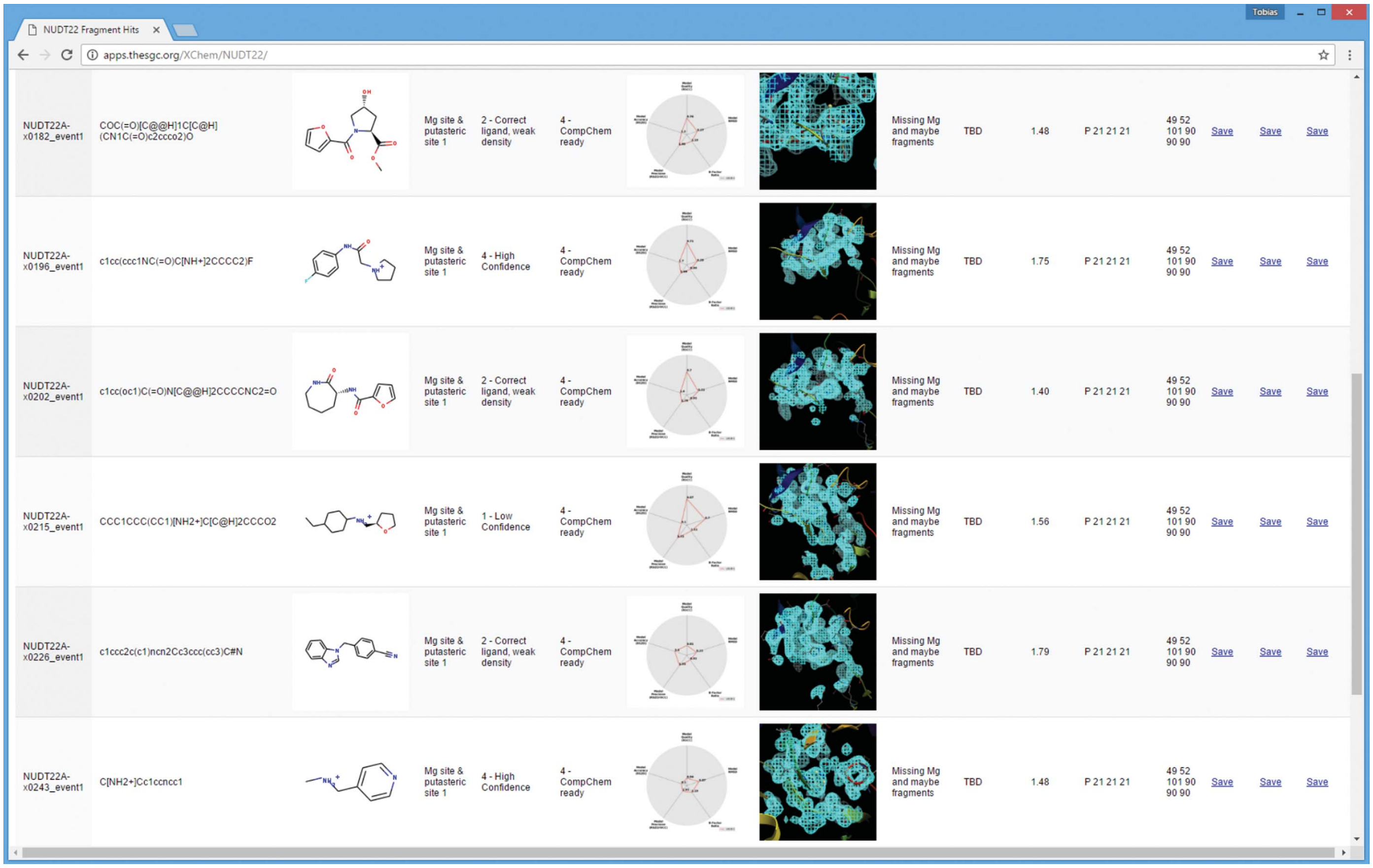
(i) Which protein–ligand structures are available?
(ii) Where is the PDB file?
(iii) Is this a ‘good model’?
(iv) Is that water/methyl/hydroxyl really where it was modelled?

XCE therefore includes a tool which converts the results in the database into an HTML document (Fig. 7). The HTML format allows the easy sharing of results either within an organization or with other scientists *via* the internet. The bound ligands are displayed in a tabular form, including details of the compound, manual annotations of reliability and links to download PDB, MTZ and CIF files. Most importantly, the recipients can themselves assess the reliability of the model: numerically by a radar plot that summarizes the reliability of the ligand and model (Pearce & von Delft, 2017) and graphically through an interactive *IcmJS* plugin (Molsoft LLC) that displays the primary evidence for the ligand, namely the *PanDDA* event map.

5. Results and discussion

XCE has been extensively tested by supporting crystallographic fragment screening at Diamond Light Source. In the course of several iterations of bug fixes, the program has become one of the cornerstones of the facility: only once the program was available in its current stable form did it become possible to completely model fragment-screening projects promptly and efficiently.

The philosophy of the program was stress-tested in a retrospective analysis of six historic fragment campaigns where the goal was to detect and model all bound ligands with the novel *PanDDA* algorithm and assemble information about how the fragment set influences the diffraction behaviour of different crystal systems. The review of 5970 individual data sets was completed by two experienced crystallographers in only 5 d, including the documenting of all experiments and finalizing all 380 protein–ligand structures. This was only possible because the program condensed huge amounts of information to make it comprehensible for the user. For instance, many data-quality statistics are available, but we confirmed that two criteria, namely the resolution limit, defined as where Mn$\langle I/\sigma(I)\rangle$ falls below 1.5, and the $R_{\rm merge}$ value for the inner resolution shell, were wholly sufficient for identifying problematic data sets. Additionally, the representation of results in tables and the possibility to sort them made it straightforward to triage data sets into those that could be taken forward and those that needed further attention. Therefore, most of the work up to manual inspection of



NUDT22A-x0182_event1	COC(=O)[C@@H]1C[C@H](CN1C(=O)c2ccco2)O		Mg site & putasteric site 1	2 - Correct ligand, weak density	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.48	P 21 21 21	49 52 101 90 90 90	Save	Save	Save
NUDT22A-x0196_event1	c1cc(ccc1NC(=O)C[NH+]2CCCC2)F		Mg site & putasteric site 1	4 - High Confidence	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.75	P 21 21 21	49 52 101 90 90 90	Save	Save	Save
NUDT22A-x0202_event1	c1cc(oc1)C(=O)N[C@@H]2CCCCNC2=O		Mg site & putasteric site 1	2 - Correct ligand, weak density	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.40	P 21 21 21	49 52 101 90 90 90	Save	Save	Save
NUDT22A-x0215_event1	CCC1CCC(CC1)[NH2+]C[C@H]2CCCO2		Mg site & putasteric site 1	1 - Low Confidence	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.56	P 21 21 21	49 52 101 90 90 90	Save	Save	Save
NUDT22A-x0226_event1	c1ccc2c(c1)ncn2Cc3ccc(cc3)C#N		Mg site & putasteric site 1	2 - Correct ligand, weak density	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.79	P 21 21 21	49 52 101 90 90 90	Save	Save	Save
NUDT22A-x0243_event1	C[NH2+]Cc1ccncc1		Mg site & putasteric site 1	4 - High Confidence	4 - CompChem ready			Missing Mg and maybe fragments	TBD	1.48	P 21 21 21	49 52 101 90 90 90	Save	Save	Save

Figure 7
Data sharing *via* HTML export. Results generated with *XCE* can be exported as an HTML document, which can be put on a web server and shared with other scientists. The figure shows the contents of the summary page; details of each modelled ligand and download links are displayed on one line. Additionally, an interactive *IcmJS* plugin (Molsoft LLC) allows display of the ligand-binding site and the respective *PanDDA* event map.

the *PanDDA* output was completed with minimal user input. Here too, only around 20% of the all data sets had to be checked for bound ligands, as they were the only ones with a statistically significant difference signal. Finally, all steps that required manipulation of models or decision making by the refiner were facilitated because all commands and annotations could be carried out within a single instance of *Coot*, allowing the users to finish tens of structures within a few hours.

While *XCE* implements all steps required to solve protein–ligand structures, the program is designed to accommodate multiple entry points, since different projects have different requirements and not all will always conform to the standard workflow. For example, while the *PanDDA* algorithm is known to be a powerful tool for hit identification, if a user decides that this is not applicable, *XCE* nevertheless supports inspection of initial electron-density maps, ligand building and refinement.

XCE depends on a relational SQLite database, but is at its heart a filesystem-based tool. While a server-based database system may have certain advantages and seem more elegant, the chosen paradigm is more transparent for the responsible scientists and allows the introduction of changes without the user having to be familiar with relational database systems. However, there are clearly some limitations to this approach, which other recently presented developments have circumvented, including *ManageCCP4i2Archive*, a tool for the collaborative sharing of large sets of structural projects undertaken across a research group, written in the *CCP4i2* structure-determination environment but using a web-based approach (M. Noble, personal communication). Both developments clearly point in the same direction and could indeed be combined with relative ease for multi-disciplinary research projects: the structures and results generated by *XCE* could be fed into *ManageCCP4i2Archive*, which could then be used to distribute the information to all the scientists involved in the project, providing a high-level platform for further structural analysis.

The algorithmic simplicity of the workflow means that most apparent errors are in fact caused by undetected crystallographic problems. The unexpected appearance of new crystal forms for a subset of the collected crystals is a typical case where the underlying problem can be quickly and easily established with *XCE*. In particular, discrepancies between the reference files provided and the data sets will almost always show up in the Maps & Restraints table when the point groups are not in agreement, the differences between unit-cell volumes are large or the $R_{\rm cryst}/R_{\rm free}$ values are high after initial refinement. The sorting mechanism allows the identity and the number of data sets affected to be rapidly identified. Additionally, more subtle differences between the data sets can be detected with a *PanDDA* data-set clustering algorithm.

Although deposition procedures at the Protein Data Bank are well established, using them to deposit dozens of structures is not realistic without helper tools. The PDB is currently developing a bulk deposition interface to encourage groups which are working on SBLD projects to contribute their structures (Stephen K. Burley, personal communication). Even so, a major bottleneck remains that all the required metadata must be collated; as *XCE* already records much of the required information, it can automatically generate valid mmCIF files. This required functionality will be fully implemented in the near future.

Acknowledgements

The authors would like to thank David Damerell, Laura Diaz Saez, Joseph Newman, Dan Pinkas and Srikannathasan Velupillai for testing of the program and helpful feedback. The SGC is a registered charity (No. 1097737) that receives funds from AbbVie, Bayer, Boehringer Ingelheim, the Canada Foundation for Innovation, the Canadian Institutes for Health Research, Genome Canada, GlaxoSmithKline, Janssen, Lilly Canada, the Novartis Research Foundation, the Ontario Ministry of Economic Development and Innovation, Pfizer, Takeda and the Wellcome Trust (092809/Z/10/Z).

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Bauman, J. D., Patel, D., Dharia, C., Fromer, M. W., Ahmed, S., Frenkel, Y., Vijayan, R. S., Eck, J. T., Ho, W. C., Das, K., Shatkin, A. J. & Arnold, E. (2013). *J. Med. Chem.* **56**, 2738–2746.

Blundell, T. L., Jhoti, H. & Abell, C. (2002). *Nature Rev. Drug. Discov.* **1**, 45–54.

Brem, J., Cain, R., Cahill, S., McDonough, M. A., Clifton, I. J., Jimenez-Castellanos, J. C., Avison, M. B., Spencer, J., Fishwick, C. W. G. & Schofield, C. J. (2016). *Nature Commun.* **7**, 12406.

Broennimann, C., Eikenberry, E. F., Henrich, B., Horisberger, R., Huelsen, G., Pohl, E., Schmitt, B., Schulze-Briese, C., Suzuki, M., Tomizaki, T., Toyokawa, H. & Wagner, A. (2006). *J. Synchrotron Rad.* **13**, 120–130.

Carvalho, A. L., Trincão, J. & Romão, M. J. (2009). *Methods Mol. Biol.* **572**, 31–56.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Danley, D. E. (2006). *Acta Cryst.* D**62**, 569–575.

Debreczeni, J. É. & Emsley, P. (2012). *Acta Cryst.* D**68**, 425–430.

Echols, N., Moriarty, N. W., Klei, H. E., Afonine, P. V., Bunkóczi, G., Headd, J. J., McCoy, A. J., Oeffner, R. D., Read, R. J., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 144–154.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Grochulski, P., Fodje, M., Labiuk, S., Gorin, J., Janzen, K. & Berg, R. (2012). *J. Struct. Funct. Genomics*, **13**, 49–55.

Hole, A. J., Baumli, S., Shao, H., Shi, S., Huang, S., Pepper, C., Fischer, P. M., Wang, S., Endicott, J. A. & Noble, M. E. (2013). *J. Med. Chem.* **56**, 660–670.

Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.

Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Cryst.* D**68**, 484–496.

Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.

Kleywegt, G. J. (2007). *Acta Cryst.* D**63**, 94–100.

Long, F., Nicholls, R. A., Emsley, P., Grazulis, S., Merkys, A., Vaitkus, A. & Murshudov, G. N. (2017). *Acta Cryst.* D**73**, 112–122.

Mooij, W. T., Hartshorn, M. J., Tickle, I. J., Sharff, A. J., Verdonk, M. L. & Jhoti, H. (2006). *ChemMedChem*, **1**, 827–838.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Structure*, **12**, 1753–1761.

Nurizzo, D., Bowler, M. W., Caserotto, H., Dobias, F., Giraud, T., Surr, J., Guichard, N., Papp, G., Guijarro, M., Mueller-Dieckmann, C., Flot, D., McSweeney, S., Cipriani, F., Theveneau, P. & Leonard, G. A. (2016). *Acta Cryst.* D**72**, 966–975.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.

Pearce, N. M., Bradley, A. R., Collins, P., Krojer, T., Nowak, R. P., Talon, R., Marsden, B. D., Kelm, S., Shi, J., Deane, C. M. & von Delft, F. (2016). *bioRxiv*, https://doi.org/10.1101/073411.

Pearce, N. M., Krojer, T. & von Delft, F. (2017). *Acta Cryst.* D**73**, 256–266.

Porebski, P. J., Cymborowski, M., Pasenkiewicz-Gierula, M. & Minor, W. (2016). *Acta Cryst.* D**72**, 266–280.

Pozharski, E., Weichenberger, C. X. & Rupp, B. (2013). *Acta Cryst.* D**69**, 150–167.

Schiebel, J. *et al.* (2016). *Structure*, **24**, 1398–1409.

Sparta, K. M., Krug, M., Heinemann, U., Mueller, U. & Weiss, M. S. (2016). *J. Appl. Cryst.* **49**, 1085–1092.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst.* D**65**, 582–601.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.

Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.

Wasserman, S. R., Koss, J. W., Sojitra, S. T., Morisco, L. L. & Burley, S. K. (2012). *Trends Pharmacol. Sci.* **33**, 261–267.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.

Winter, G. & McAuley, K. E. (2011). *Methods*, **55**, 81–93.