US National Committee for Crystallography

Exploring Structural Database Use in Crystallography: Workshop Series

The U.S. National Committee for Crystallography (USNC/Cr) of the National Academies of Sciences, Engineering, and Medicine hosted an online workshop series for students and researchers on the use, development, and maintenance of crystallographic and structural databases. Encompassing macromolecular, small molecule, and powder diffraction information, the series will include 11 modules each meeting for one or two days:

Opening Keynote - John Helliwell - The Exemplary Crystallography and Structural Databases
ICSD Searching and using the Inorganic Crystal Structure Database (2 days)
ICDD The International Center for Diffraction Data Powder Diffraction File™: Database Concepts And Applications (2 days)
PDB Introducing the Protein Data Bank: 3D Macromolecular Structure Data Deposition, Validation, Biocuration, Archiving, and Delivery for Researchers, Educators, and Students Worldwide (2 days)
EMDB Introducing the Electron Microscopy Data Bank (EMDB) -- a public repository for electron cryo-microscopy volume maps and tomograms of macromolecular complexes and subcellular structures
SBGrid Using the SBGrid Software Installer, AppCiter, and SBGrid Data Bank on PCs, HPC clusters and AWS cloud
COD Depositing and managing data in the Crystallography Open Database (COD)
IRRMC Access to experimental data is crucial for scientific reproducibility (Integrated Resource for Reproducibility in Macromolecular Crystallography)
COD Searching and getting data from the Crystallography Open Database(COD)
CSD Working with the Cambridge Structural Database (CSD): searching, analyzing and depositing data using Cambridge Crystallographic Data Centre (CCDC) tools (2 days)
COD Under the hood: Building your own scientific database

Graduate students, postdoctoral fellows, faculty members and researchers in any of the crystallographic, diffraction, and imaging sciences affiliated with the International Union of Crystallography (IUCr) are encouraged to register and participate in the training sessions that interest them.

Registration for this online series was free, but is no longer available. However, recordings of the sessions are available. The workshop seris main page is:
https://www.nationalacademies.org/our-work/exploring-structural-database-use-in-crystallography-a-usnccr-workshop-series

Exploring Structural Database Use in Crystallography: A USNC/Cr Workshop Series

2022 Workshop at a Glance

	Sun	Mon	Tue	Wed	Thur	Fri	Sat
MARCH	20	**21 - Keynote** John Helliwell 11a-12:15p EDT Video recording	22	**23 - ICSD** Inorganic Crystal Structure Day 1 11a-1p EDT Video recording	**24 - ICSD** Inorganic Crystal Structure Day 2 11a-1p EDT Video recording	25	26
	27	**28 - ICDD** Powder Diffraction File Day 1 11a-2p EDT Video recording	**29 - PDB** Macromolecular Structure Day 1 11a-1p EDT Video recording	**30 - ICDD** Powder Diffraction File Day 2 11a-2p EDT Video recording	**31- PDB** Macromolecular Structure - Day 2 11a-1p EDT Video recording	1	2
APRIL	3	**4 - EMDB** Electron Microscopy DB 11a-2p EDT Video recording	**5 -SBGrid** Structural Biol Software and DataBank 11a-2p EDT Video recording	**6 - COD** Deposit and Manage Cryst. Open DB data 10a-2p EDT Video recording	**7 - IRRMC** Protein Crystal Raw data archive 11a-2p EDT Video recording	8	9
	10	**11 - COD** Search Cryst. Open DB 10a-2p EDT Video recording	**12 - CSD** Cambridge Structure DB Day 1 11a-2p EDT Video recording	**13 - CSD** Cambridge Structure DB Day 2 11a-2p EDT Video recording	**14 - COD** Building your own Scientific Database 10a-2p EDT Video recording	15	16

The workshop was sponsored by the National Institute of Standards and Technology (NIST).

For more information please contact Ana Ferreras, Email: aferreras@nas.edu

Workshop website
https://www.nationalacademies.org/our-work/exploring-structural-database-use-in-crystallography-a-usnccr-workshop-series

Video Recording Showcase
https://vimeo.com/showcase/9403870

Keynote: The Exemplary Crystallography and Structural Databases

Mar 21, 2022 11a-12:15p EDT (16-17:15 CET)

Crystallographers and structural scientists have long accepted the maxim "*Take nobody's word for it*" because "*the science is in the data*". Indeed, from the outset in the first crystal structure report (Bragg 1913), the diffraction data were included. Moreover, they have been among the first scientific communities to exploit technology developments for expanding their data archiving scope. The International Science Council accords great importance to data and has a dedicated committee on data, known simply as CODATA, with many decades of experience of good policy and practice. Likewise, the National Academies of Sciences, Engineering, and Medicine reports include a cogent, recent, description of the reproducibility and replicability of science within which scientific data are a solid foundation. This excellent educational course will describe the diverse and thriving ecosystem of crystallographic and structural databases. This ecosystem is underpinned by data exchange standards, the crystallographic information framework "*cif* ", and the desire to ensure the highest achievable quality, making a strong world-wide data infrastructure. There is then an exemplary provision for molecular structure scientists of collections of their data in these databases, be they for biology, chemistry or materials science, which is widely admired. Indeed, the use of these vast collections of data is extremely broad, and yield strong basic science and high societal impacts, which the course contributors will each describe.

Speaker: John R Helliwell, Emeritus Professor of Chemistry, University of Manchester, UK and DSc in Physics, University of York, UK; International Union of Crystallography Chairman of the Committee on Data and its Representative to CODATA.

Bragg, W.L. (1913) *The structure of some crystals as indicated by their diffraction of X-rays* Proc. R. Soc. London, Ser. A89, 248–277.

National Academies of Sciences, Engineering, and Medicine 2019. *Reproducibility and Replicability in Science.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

ICSD Mar 23 & Mar 24, 2022

The Inorganic Crystal Structure Database (ICSD) is the world's largest database for fully determined inorganic crystal structures. It is made available to the scientific community and industry by FIZ Karlsruhe. ICSD contains the crystallographic data of all published crystalline inorganic structures, including atom coordinates, dating back to 1913. Organometallic and theoretical structures have been added within the past years. The ICSD data are of excellent quality. Only data that have passed thorough quality checks are included. As the world's leading provider of scientific information on inorganic crystal structures, we take full responsibility for database production, maintenance and quality control, and we ensure that the ICSD database and our software solutions meet the highest possible quality standards.

In this 2 part tutorial, we will give a detailed overview of the search masks in ICSD and explain tips and tricks on how to best use them. We will also give an overview of the different ways to search in ICSD - from simple searches to more complex searches with the tools for combining searches in ICSD or with the "Expert Search" command line tool.

The tutorial session will include some questions that participants can work on using the ICSD. We will be around to give tips or further help if needed. Of course, we will also be happy to help with problems posed by participants.

Both parts are designed for beginners, but will also include some advanced tips and tricks for more familiar users of the database. There will also be plenty of time to ask questions about your specific challenges.

2-day series: Inorganic Crystal Structure Database

Learning Objectives

1. Participants will be provided an overview of the search masks in ICSD including tips and tricks on how to best use them.
2. Participants will learn different ways to search in ICSD - from simple searches to more complex searches with the tools for combining searches in ICSD or with the "Expert Search" command line tool.

Take Away Skills

1. Ability to use the ICSD to answer interesting problems as modeled through practice problems.

ICDD Mar 28 & Mar 30, 2022

Crystallographic databases play a vital role in materials research, influencing materials development and providing a reference for materials characterization. Design, data curation, and data management are all critical factors in developing a successful and useful database. This presentation will (1) focus on the processes used in creating the Powder Diffraction File (PDF®) including quality, reliability, management and accessibility of data and (2) highlight materials analysis applications.

The International Centre for Diffraction Data (ICDD®) Powder Diffraction File (PDF®) is a powerful database for materials characterization that has been used extensively by the scientific community. Starting with 1000 cards in 1941, the database has grown to contain over 1 million unique material data sets. The Powder Diffraction File has a wealth of information that a materials scientist can take advantage of in various ways, from materials identification, characterization, computation to design. The Powder Diffraction File in Relational Database Format (RDB) format contains extensive chemical, physical, bibliographic and crystallographic data including atomic coordinates enabling characterization and computational analysis.

Proper database structure, data validations and phase-type classifications are crucial in making any database useful and reliable. Various structural and chemical classifications implemented in the database will be presented in detail. These classifications are important in data mining studies and optimizing diffraction pattern search/match methods. While using a database, it is important to know the quality of the crystal structure, diffraction pattern data and any data field of interest found in the database. With the varying quality of published data in the literature, the PDF database editorial review processes require rigorous data evaluation methods to define data based on its quality. This critical evaluation is the rate determining step in populating a curated PDF database. Once the database has been established it is important to explore how the database can best be applied to solve problems in materials science. In the second part of this workshop, various database applications from phase identification, quantitative analysis and materials characterization using datamining applications will be presented. Advanced features in the PDF including atomic coordinates for Rietveld refinement techniques; amorphous and nano material references; digital simulation tools for evaluating X-ray, synchrotron, electron and neutron diffraction data as well as crystallite size and analysis of two-dimensional diffraction data will also be discussed

2-day Series: The ICDD® Powder Diffraction File™: Database Concepts And Applications

Learning Objectives

1. Participants will learn about the processes used in creating the Powder Diffraction File (PDF®) including quality, reliability, management and accessibility of data.
2. Participants will learn how to use the ICDD for materials analysis applications.Participants will learn various database applications from phase identification, quantitative analysis and materials characterization using datamining applications.
3. Participants will be exposed to advanced features in the PDF including atomic coordinates for Rietveld refinement techniques, amorphous and nano material references, digital simulation tools for evaluating X-ray, synchrotron, electron and neutron diffraction data, and crystallite size and analysis of two-dimensional diffraction data.

Take Away Skills

1. Ability to use the ICDD for materials analysis applications
2. Ability to use multiple database applications including advanced features.

PDB Mar 29 & Mar 31, 2022

The Protein Data Bank (PDB; https://www.wwpdb.org) is the single global archive for preserving and disseminating information about the experimentally-determined three-dimensional (3D) structures/shapes of proteins, nucleic acids, and complex assemblies. Structural biologists working on every inhabited continent have contributed more than 185,000 experimentally determined structures to the PDB since it was established in 1971 as the first open-access digital data resource in biology. PDB data inform our understanding of fundamental biology, biomedicine, bioenergy, and bioengineering/biotechnology. The information is used by many millions of basic and applied researchers, educators, and students working and learning in every sovereign country recognized by the United Nations. Since 2003, the archive has been managed by the Worldwide Protein Data Bank (wwPDB) partnership, which includes PDB data centers in the US (RCSB PDB; https://www.rcsb.org), UK/Europe (PDBe; https://www.pdbe.org), and Asia (PDBj; https://www.pdbj.org) plus two specialty data archives for electron microscopy (EMDB; https://www.emdb.org) and NMR spectroscopy (BMRB; https://www.bmrb.io). This two-day workshop, sponsored by the RCSB Protein Data Bank (US PDB data center), will introduce participants to the science and technology of 3D macromolecular structure data deposition, validation, biocuration, archiving, and delivery.

2-day Series: Introducing the Protein Data Bank: 3D Macromolecular Structure Data Deposition, Validation, Biocuration, Archiving, and Delivery for Researchers, Educators, and Students Worldwide

Learning Objectives

1. Appreciate the importance of the PDBx/mmCIF data standard that underpins organization of all PDB archival data
2. Access and appropriately use the wwPDB OneDep system for macromolecular structure deposition
3. Understand how small molecule data in the PDB are organized and accessed
4. Evaluate the quality and accuracy of a PDB structure using the wwPDB validation report
5. Understand how to use the Mol* molecular viewer to visualize PDB structures
6. Appreciate wwPDB plans for establishing a Next Generation PDB Archive
7. Participants will learn about the science and technology of 3D macromolecular structure data deposition, validation, biocuration, archiving, and delivery.

EMDB Apr 4, 2022

Background

The Electron Microscopy Data Bank (EMDB) is a public repository for electron cryo-microscopy volume maps and tomograms of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography, and electron crystallography

Introducing the Electron Microscopy Data Bank (EMDB) -- a public repository for electron cryo-microscopy volume maps and tomograms of macromolecular complexes and subcellular structures.

Learning Objectives:

1. Participants will come away with knowledge of the different methods used to produce high resolution EM structures.
2. Participants will learn how to use the EMDB website to browse for EM structures of interest.
3. Participants will be introduced to validation metrics that allow them to assess the quality of EM data both in the EMDB archive and more generally.
4. Participants will be introduced to EM visualization software.
5. Participants will be briefly introduced to EMPIAR and AlphaFold DB.
6. Participants will embark on interactive, assisted, exercises to learn and reinforce their understanding of EM and the EMDB.

Take Away Skills:

1. Ability to search the EMDB archive
2. Ability to interpret the validation metrics of entries in the EMDB archive
3. Ability to visualize EM data

SBGrid Consortium Apr 5, 2022

SBGrid (www.sbgrid.org) is a consortium of 427 structural biology groups that utilize homogeneous stacks of scientific applications. SBGrid "Factory" at Harvard Medical School, which is led by Dr. Sliz actively curates over 1000 software titles ranging from tools used in crystallography, electron microscopy, computational chemistry to computational biology (biogrids.org). Multiple versions of all applications, the supporting libraries, and configuration files are maintained. The SBGrid consortium members use the extensive software collection, and work with the Factory to improve the quality SBGrid environment.

With software support activities dating back to 2001, the SBGrid collection is the largest global library of ready-to-execute scientific software. As part of its activities, SBGrid also maintains training materials (https://www.youtube.com/user/SBGridTV), software metadata, scientific source code and information about software citations, which is available through the SBGrid AppCiter (https://sbgrid.org/software/). More recently, we also developed a data bank to archive and validate X-ray diffraction, MicroED and LLSM datasets from SBGrid laboratories (data.sbgrid.org).

Using the SBGrid Software Installer, AppCiter, and SBGrid Data Bank on PCs, HPC clusters and AWS cloud.

Learning Objectives

- *Learn how to use SBGrid Software Installer, AppCiter, and SBGrid Data Bank*
- *Learn how to deploy SBGrid applications and experimental data on personal computers, HPC clusters, and in AWS.*
- *Learn how to effectively utilize SBGrid applications in support of CryoEM, small-molecule docking and structure prediction.*
- *Scientific software curation process - skills and expertise required to compile scientific software.*

Take Away Skills

- *Ability to effectively utilize SBGrid open resources (SBGrid Data Bank, SBGrid TV, AppCiter, Scientific Software Database)*
- *Ability to effectively utilize consortium-supported software (SBGrid and BioGrids stacks of applications), and to contribute to SBGrid*
- *Basic understanding of software curation process and setting computing environment in AWS.*

COD Apr 6, 2022

(Crystallography Open Database) and TCOD (Theoretical Crystallography Open Database)

Using an open-access distribution model, the Crystallography Open Database (COD, http://www.crystallography.net) collects all known ‘small molecule / small to medium sized unit cell’ crystal structures and makes them available freely on the Internet. As of today, the COD has aggregated over 480,000 structures, offering basic search capabilities and the possibility to download the whole database, or parts thereof using a variety of standard open communication protocols. A website provides capabilities for all registered users to deposit published and so far unpublished structures as personal communications or pre-publication depositions. Such a setup enables extension of the COD database by many users simultaneously. This increases the possibilities for growth of the COD database, and is the first step towards establishing a world wide Internet-based collaborative platform dedicated to the collection and curation of structural knowledge.

Depositing and managing data in the Crystallography Open Database (COD) and Theoretical Crystallography Open Database (TCOD)

Learning Objectives

1. Participants will learn how to deposit their data into the database including what information is needed and how to deposit raw data.
2. Participants will learn how to maintain their data records in COD and TCOD.
3. Participants will be informed of quality standards such as: how personal data is used, how on-hold records are released to the public, and how records are peer reviewed.

IRRMC Apr 7, 2022

The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) includes a repository system and website designed to make the raw data of protein crystallography more widely available. Its focus is on identifying, cataloging and providing the metadata related to datasets, which could be used to reprocess the original diffraction data. The intent behind this project is to make the resulting three dimensional structures more reproducible and easier to modify and improve as processing methods advance.

Access to experimental data is crucial for scientific reproducibility

Learning Objectives

1. Participants will learn about the proteindiffraction.org repository of diffraction experiments used to determine structures that are din the protein data bank
2. Learn about the IRRMC tools used for working with data from the archive
3. Learn about progress in using neural networks to find anomalies in diffraction images and malfunctions of experimental systems
4. Learn about wwPDB plans for archival of integrative structures, whose models rely on combining information from multiple techniques and sources. The experimental data supporting the models may come from techniques such as X-ray crystallography (X-ray), nuclear magnetic resonance (NMR) spectroscopy, three-dimensional electron microscopy (3DEM), small-angle solution scattering (SAS), chemical cross-linking mass spectrometry (CX-MS), Förster resonance energy transfer (FRET), electron paramagnetic resonance spectroscopy (EPR), hydrogen-deuterium exchange mass spectrometry (HDX-MS), and other biophysical and proteomics methods.
5. Learn how access to unprocessed experimental data impacts scientific reproducibility.

Speakers:

John Helliwell, Emeritus Professor of Chemistry, University of Manchester, UK and DSc in Physics, University of York, UK; International Union of Crystallography Chairman of the Committee on Data and its Representative to CODATA.

David Cooper, Ph.D., Assistant Professor of Research at the University of Virginia

Dariusz Brzezinski, Ph.D., D.Sc., Associate Professor at the Institute of Computing Science, Poznan University of Technology

Brinda Vallat, Ph.D., Assistant Research Professor at the Institute for Quantitative Biomedicine at Rutgers and is a member of the RCSB Protein Data Bank

Wladek Minor, Ph.D., Harrison Distinguished Professor of Molecular Physiology and Biological Physics at the University of Virginia.

COD Apr 11, 2022

(Crystallography Open Database) and TCOD (Theoretical Crystallography Open Database)

Using an open-access distribution model, the Crystallography Open Database (COD, http://www.crystallography.net) collects all known ‘small molecule / small to medium sized unit cell’ crystal structures and makes them available freely on the Internet. As of today, the COD has aggregated over 480,000 structures, offering basic search capabilities and the possibility to download the whole database, or parts thereof using a variety of standard open communication protocols. A website provides capabilities for all registered users to deposit published and so far unpublished structures as personal communications or pre-publication depositions. Such a setup enables extension of the COD database by many users simultaneously. This increases the possibilities for growth of the COD database, and is the first step towards establishing a world wide Internet-based collaborative platform dedicated to the collection and curation of structural knowledge.

Searching and getting data from the Crystallography Open Database (COD) and Theoretical Crystallography Open Database (TCOD)

Learning Objectives

1. Participants will gain a basic understanding of the structure of the COD including contents, scope, versioning, curation and identification principles of the COD.
2. Participants will understand how computations and queries are performed in COD.
3. Participants will learn how COD integrates with other databases.

Take Away Skills

1. Ability to use the basic computations and queries of COD

CSD Apr 12 & Apr 13, 2022

Background

The Cambridge Crystallographic Data Centre (CCDC) is the curator of the Cambridge Structural Database (CSD), the world's largest repository of fully curated organic and organometallic experimental crystal structures. The large amount of reliable data makes the CSD an excellent place to gather insights into chemical and structural space.

2-day series: Working with the Cambridge Structural Database (CSD): searching, analyzing and depositing data using Cambridge Crystallographic Data Centre (CCDC) tools

Learning Objectives

1. Participants will explore how the CCDC's tools allow users to search for structures of interest.
2. Participants will learn how to generate knowledge from statistical trends.
3. Participants will understand how to submit their own data to the database.

Take Away Skills

1. Ability to search structures and statistical trends within the database
2. Ability to submit their own data to the database

*Note: Graphical user interfaces and Python-based scripting will both be demonstrated.

Building your own database Apr 14, 2022

Targeted to anyone wishing to manage their (own) data reliably in a long term,
this session will presented by the team from the Crystallography Open Database (COD).
Some of the items to be touched on include:

- Contents and scope of the databases (once again; here you need to choose it yourself ;))
- Sensitive personal and medical data – a disclaimer (we will not use such data in the demo ;))
- Unique, stable identifiers. Stable identifier systems: ARK, Handle, DOIs. A rant about DOIs.
- Data versioning; version control systems (Subversion, CVS, RCS, Git)
- Data formats (CIF, STAR, XML, Yaml, JSON, CSV)
- Low-level data encoding; Unicode; encoding transparency
- Data dictionaries
- The use of relational model
- Relational databases; their use as fas search cache; data flow
- NoSQL databases (short mention; we do not use them a lot)
- Data curation principles (do not invent data; consult sources; document changes and *why* you made them; maintain data provenance; maintain metadata)
- Web access to your data: REST, HTTP, stateless protocols, paging
- Web security issues: XXS, injections, updates, isolation (virtual machines, Docker, etc.)

Under the hood: rolling out your own scientific database

Learning Objectives

1. Participants will learn the basics of preparing their own database including choosing stable identifiers, versioning, using sensitive data, the use of relational models, and data curation.
2. Participants will learn the nuts and bolts of databases such as data formats, low-level data encoding, and data dictionaries.
3. Participants will learn logistical issues such as web access and web security.

Take Away Skills

1. Ability to work toward the preparation of a database with awareness of potential issues that must be addressed