New Frontiers beyond one million: new horizons for structural chemistry

ACS August 2019, San Diego

Dr. Juergen Harter
CEO
28th August 2019
Running order and key themes

• Data growth
  • CSD growth / structural chemistry data growth
  • Chemical space, diversity, coverage, trends over time

• Advances in instrumentation and techniques
  • Yielding more data with higher accuracy and precision

• Automation & Software development
  • Evolution of existing tool set / future tools development

• Where does the community best fit?
  • Publishing advances

• Interdisciplinarity – connecting / linking other domains
  • Biology / Pharmacology / Medicine

• Insights
  • AI / ML based
  • From advanced search & visualisations
  • Derived from underlying high-quality data

• New science & innovation
  • Powered by scientific hypotheses underpinned by data
  • Spotting emerging trends
  • Digital Drug Design & Manufacturing Centres

• Semantic representation & analyses
  • Machine / human experts - symbiosis
  • Reasoning / Inferencing (based on ontologies)

• Quality and precision of data and meta data
  • Evolution of data standards, and data architecture (more capabilities & flexibility)
  • Data initiatives (FAIR, InChI, PIDs, CIF2.0)
  • Bulk operations: integrity checks, quality analyses, data metrics
The Cambridge Structural Database (CSD)

- Over 1 Million small-molecule crystal structures
- Over 80,000 datasets deposited annually
- Structures available for anyone to download
- Links to over 1,000 journals
- Enriched and annotated by experts
- Access to data and knowledge

Every published structure:
- Inc. ASAP & early view
- CSD Communications
- Patents
- University repositories

Every entry enriched and annotated by experts
Discoverability of data and knowledge
Sustainable for over 54 years

One of 1st – METALD Metaldehyde – published 1936
250,000th – IBEZUK Conducting metal-dinitrogen polymer
500,000th – EFEMUX01 Lamotrigine – an anti-convulsant drug
750,000th – ZOYBIA Co-crystal of vanillic acid and theophylline
1,000,000th – XOPCAJ Chalcogen bonding catalyst
Inside the CSD

Organic
- 43%
- At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po.

Metal-Organic
- 57%
- Single Component: 56%
- Multi Component: 44%
- Not Polymeric: 89%
- Polymeric: 11%

Organic
- Drugs
- Agrochemicals
- Pigments
- Explosives
- Protein ligands

Metal-Organic
- Metal Organic Frameworks
- Models for new catalysts
- Porous frameworks for gas storage
- Fundamental chemical bonding

Additional data
- 10,860 polymorph families
- 169,218 melting points
- 840,667 crystal colours
- 700,002 crystal shapes
- 23,622 bioactivity details
- 9,740 natural source data
- > 250,000 oxidation states

Links/subsets
- Drugbank
- Druglike
- MOFs
- PDB ligands
- PubChem
- ChemSpider
- Pesticides
From Experiments to Knowledge

The aggregation of experimental datasets provides a foundation for resources that enable structural knowledge to be applied to scientific challenges across sectors and domains.

**CSD-Data:** Full CSD data access or layered/sliced high quality data access for AI/ML operations

**CSD-System:** Find (advance search), analyse and communicate crystal structures

**CSD-Materials:** Behaviour and properties of new materials

**CSD-Discovery:** Protein and ligand-based design of new drugs

**CSD-Enterprise:** Seamless & integrated platform and orchestrated tools for interactive drug design & drug manufacture (design centres)

Association of chemistry and crystallography is key for enabling discovery of new insights.
Instrumentation & Techniques

MicroED – total sample characterisation

Better Robotics & AI for Sample Characterisation – Find all the polymorphs?

X-Ray Free Election Lasers – Experimental Molecular Dynamics

Miniaturization & Automation of Routine X-Ray Structure Solution – Crystallography for the Bench Chemist

Experimental characterisation in concert – the sum is greater than the parts

Moores’ Law & Fast Crystal Structure Prediction
The Opportunities & Challenges for CCDC

• More structures, and much more associated data 😊
  • Automation of curation at source
  • Richer data sets: more experimental information

• But more capability could lead to less sharing
  • Less recognition that each structure is still valuable ...

• Quality & Trust
  • What’s experimental? What’s theoretical?
  • Less investment of expertise; what’s fit for purpose
  • Divergence of techniques: education of data consumers needed
Software / Automation / Workflows

- Code / software development advances
  - Replacing old legacy code
  - Agile software development
- Software architecture progress
  - Modern cloud architecture, improved UX/UI
- Python API
  - Captures 90% of our functionality of the present CCDC software suites
  - Enabling interoperability with other platforms and software tools
- Maintain and evolve various software tools
  - checkCIF & others
- Advances in curation and in validation
- Automation Automation Automation!!!
  - Workflow platforms (KNIME, Biovia Pipeline Pilot, Orion etc.)
Product roadmap: Overarching product themes

- **User experience**
  - Consolidation and rationalisation of user interfaces to ensure a clear, coherent suite of products for our user communities. Emphasis on user experience, streamlining key activities in the interfaces to make CSD products more straightforward and intuitive.

- **Data management**
  - Effective management of user's structural data alongside the world's published structural data (the CSD) to ensure seamless search & analysis across the combined structural knowledge.

- **Visualisation**
  - High quality 3D visualisation is key to many aspects of structural science and this will be a strong product theme into the future.

- **Advanced search**
  - Powerful and intuitive searching across 0D, 1D, 2D and 3D aspects of structural data allowing exploration and flexible ideas generation.

- **Molecular conformation**
  - The CSD is a key source of experimental molecular geometry knowledge and provides both effective assessment of geometries and prediction of likely geometries.

- **Intermolecular interactions**
  - Molecular recognition is critical in drug design, crystal engineering and many other areas of structural science. The CSD is central to the world's in-depth understanding of 3D intermolecular interaction geometries.

- **Structural property insights**
  - Combine the highest quality structural data with cutting edge machine learning approaches to deliver valuable insights. Predictive analytics to help decision-making and innovation across all areas of structural science.
Community

- How will they want to work over the next decade?
- Comprehensivity
  - Published vs. unpublished data
- Deposition tools (making it easier)
- Software tools
- Validation tools (checkCIF, Platon, Squeeze)
  - CIF 2.0? / change in community standard hard
  - X-ray powder diffraction – would need experts stepping forward to incorporate in future improved checkCIF (presently more for small molecules)
- GitHub – management of model stores / algorithms (sets of)
- User Forums
  - Exchanging experiences
  - Frequently asked questions
The FAIR Data Principles

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons


DATA SHOULD BE

Findable
Interoperable
Accessible
Reusable

BY HUMANS AND MACHINES

From https://www.slideshare.net/ConfOA/workshop-fair-data-principles
Making the CSD even FAIRer

• Greater interlinking
  • Importance of standard identifiers
  • Creating Knowledge Graphs

• Machine accessibility
  • Importance of APIs – e.g. CSD Python API
  • Machine accessibility through web – we don’t have this but probably should

• Richer representations
  • Exploiting semantic technologies to achieve greater interoperability
  • Going beyond diffraction data to other experimental and computational results
  • For example, setting technical and scientific standards for Crystal Structure Prediction (CSP)

• FAIR vs Open vs Sustainability
Interoperability and Standard Identifiers

- **DOIs for Digital Objects**
  Other persistent identifiers are available (ARks, Handles, etc.)

- **ORCID iDs for Researchers**
  30% of current CSD depositors provide an ORCID iD

- **Research Organization Registry for Institutions**
  See activities of the Organization Identifier Working Group

- **InChIs for Chemical Structures**
  IUPAC

- **Identifiers for earths science samples and specimens**

- **Identifiers for antibodies, organisms, cell lines, etc.**
Ontologies: Everything’s a graph really... PID Graph / Knowledge Graph / Open Science Graph

Project FREYA
Connected Open Identifiers for Discovery, Access and Use of Research Resources

See also RDA Interest Group – Open Science Graphs for FAIR Data https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig

PID graph – research objects connected by PIDs


https://lod-cloud.net/
An article with a title
Author, A., Author B, AC

Contact email

Introduction
Here is some introduction about the article. There will be lots of references. I should just copy and paste some text instead of writing things.

Methods
Here would be the description of how the data was collected. Crystallographic information would be included here, probably instrumentation information which hopefully will also be included in the CIFs.

PID Graph
Each circle represents something which can be identified by a PID, lines indicate connections between PIDs.

Image adapted from https://doi.org/10.5438/jwvf-8a66 by Natalie Johnson, CCDC
Chemical and Biological Interoperability

• Interdisciplinarity
  • PDB (wwwPDB), ICSD/FIZ, ICDD etc.
  • From small and simple molecules to big and complex ones – covering them all

• Interoperability
  • interoperability between the two fields of chemical crystallography and protein crystallography is difficult
  • Advanced searching – UI/UX seamless

• Precision
  • PDB, ligands, CrossMiner – protein forum, modified proteins (storage)

• Linking biology & chemistry
  • Drug development harnessing the CSD and PDB
  • adding pharmacology, medicine, and other domains reliant on structural chemistry

• Impact of CSD
  • Our high quality curation methods
Why data quality matters greatly!

If Your Data Is Bad, Your Machine Learning Tools Are Useless
by Thomas C. Redman
APRIL 02, 2018

Want to use AI and machine learning? You need the right infrastructure

IT is being tasked with supporting artificial intelligence and machine learning initiatives, and that requires thinking broadly about infrastructure needs today and tomorrow.

The right infrastructure, quality data needed


“As the director of datamine decision support systems, I’ve delivered more than 80 data-intensive projects across several industries and high-profile corporations. These include data warehousing, data integration, business intelligence, content performance, and predictive models. In most cases, data quality proved to be a critical factor for the success of the project.”

CCDC as Data Trust

CCDC as a Trusted Data Repository

Data Trustworthiness

Data Quality

High Standards

High Precision

Wealth of Data, Information and Knowledge

Gold Sets and Subsets, for Machine Learning and Algorithm Testing

CSD Inside...
Using AI and machine learning

• AI and machine learning techniques are evolving rapidly

• But the consequences of using poor quality data can be far reaching
  • Incorrect scientific conclusions
  • Wasted investment and effort
  • A loss of trust
  • Ultimately poor business or research decisions

Many of the most pressing challenges facing AI today resolve around its poor-quality training data......while algorithmic improvements could help, so could having proper training data.

What could CCDC do going forwards to facilitate the CSD in Machine Learning?

1. Community Model Stores for CSD derived models
2. Additional Curation of Critical Data
3. Better APIs for access to data in easily consumable forms
4. Cross-association with reliable meta-data
Using the collection to build predictive models

1. Select diverse structures,
2. Calculate properties using CPU intensive methods
3. Use machine learning to build predictive models of the property

Taken from Federico M. Paruzzo et al, Nature Communications 9, 4501 (2018)
Using the collection to find new ideas

Virtual screening for organic high-mobility semiconductors

The challenge:
- Find new organic semi-conductors

The method:
- Screen the CSD for existing structures
- Calculate key parameters using high-computational cost modelling

Christoph Schober; Karsten Reuter; Harald Oberhofer; J. Phys. Chem. Lett. 2016, 7, 3973-3977. DOI: 10.1021/acs.jpclett.6b01657
Virtual screening for organic semiconductors

Published in: Christoph Schober; Karsten Reuter; Harald Oberhofer; J. Phys. Chem. Lett. 2016, 7, 3973-3977. DOI: 10.1021/acs.jpclett.6b01657 Copyright © 2016 American Chemical Society
Using the collection to help predict structures

CCDC Blind Test Showcases Major Advance in Crystal Structure Prediction Methods

- November 03, 2015

The Cambridge Crystallographic Data Centre (CCDC) announces that the results of its 6th blind test of crystal structure prediction methods demonstrate significant advancement in crystal structure prediction methods in comparison with previous tests. This year, structures of all of the test systems, which included the generation of polymorphs, salts and hydrates, were generated by one or more methods. In addition, experimental structures were predicted to be the most stable form.

CRYSTAL CHALLENGE

The 3D structure that a molecule adopts in a crystal is very difficult to predict — but defines what properties the molecule has.

The structural formula of a molecule reveals which atoms are connected at a 2D level.

Chemists are making progress at predicting how complex molecules will assemble in 3D space — there are millions of possibilities.

The 3D orientation repeats in a crystalline lattice with a structure that dictates the molecule’s mechanical, chemical and physical properties.

Software predicts slew of fiendish crystal structures

Chemists succeed at forecasting how complex molecules will assemble in 3D.

Elizabeth Gibney

CCDC
CSP Blind Test VII – coming soon...

• A way to benchmark the predictive methods for CSP
  • The blind test is true methods validation; a hard challenge
• Help drive on methods development in CSP
  • Encourage new ways of tackling the problem
• Help promote the methods going forwards
• Help to inform the user community on the state-of-the-art
Several putative forms exist within similar energy (possibly to a known form) – polymorphic red flag

Nature of unobserved forms
CSP in the year 2119

- Identification of only the forms that can be made
- Predicted properties in solution
- Predicted particle properties
- Creation recipes at ambient and non ambient conditions
- Stability predictions and rankings

REPORT

- Identification of only the forms that can be made
- Predicted properties in solution
- Predicted particle properties
- Creation recipes at ambient and non ambient conditions
- Stability predictions and rankings
Advancing Science at CCDC

Insights from data-driven prediction
- Prediction of protein binding
- Prediction of crystalline particle properties
- Prediction of crystalline bulk properties
- Control of molecular conformation
- Prediction of molecular properties

Insights from analogous data
- Analysis of the solid form
- Analysis of structural properties
- Interpretation of protein binding
- Ideas generation for molecular design
- Understanding of molecular conformation
- Understanding molecular interactions

Current Target Areas

- Small Molecule Crystallography
- Protein Crystallography
- Compound Discovery in Pharmaceuticals & Agrochemicals
- Compound Development in Pharmaceuticals & Agrochemicals
- Non-pharma Materials Design
- Early stage discovery and optimisation
- Materials/Product development
- Catalysis, MOFs

Structural validation
- Interpretation
- Visualization
- Structural comparison
- Identification
- Orientation & Deformation
- Fracture prevention
- Powder diffraction
- Structural refinement
- Structural validation
- Structural model building
- Catalysis
- Nanotechnology design
- Carbon capture
- Hydrogen storage
- Molecular electronics
- Virtual screening of new molecular entities
- Ideas for hit compound optimisation
- Interpretation of protein-ligand binding
- Relationships between protein structures
- Validation of compounds
- Pharmacophore creation
- Target validation
- Informatics-driven risk assessment
- Ideas for formulation
- Interpreting crystal structure prediction
- Crystal structure prediction
- Solubility prediction & assessment
What could we do if ...

- Every (organic) structure in the CSD was annotated with, say, its full electrostatic potential surface, calculated from a high level of theory?
  - Better understanding of structural properties and trends?
  - Better forcefields for the solid state?
  - Better predictive models?
- What other properties might empower more research?
- CPU availability means these ideas are more tractable now
Visualisation and VR / AR

- New drug and material discoveries to be untangled in VR (Bristol)

- Presented by Marc Baaden of CNRS at IUPAC:
  - Visualisation using gaming engines
    - [http://www.baaden.ibpc.fr/umol/](http://www.baaden.ibpc.fr/umol/)
  - MinOmics, an Integrative and Immersive Tool for Multi-Omics Analysis
    - [https://doi.org/10.1515/jib-2018-0006](https://doi.org/10.1515/jib-2018-0006)
  - BioSpring – Interactive Molecular Dynamics
    - [https://sourceforge.net/projects/biospring/](https://sourceforge.net/projects/biospring/)
Digital First

Digital Drug Design & Manufacturing Centres

- Interactive knowledge-based precision drug design & manufacturing, data-driven and real-time
- Using knowledge-based tools to evaluate solid-form design and risk assessment
- A drug designer answers questions, knowledge-driven by the CSD and its tools

Material Sciences ➔ Discovery Sciences ➔ Medicinal Chemists ➔ Computational Chemists

Chemistry ➔ Biology

Software Tools, Automation, AI/ML, Workflows

Structural Chemistry Data

- High quality and precision, highest standard = Data Trust
- Comprehensive coverage, high level of interlinkedness, excellent meta data
New Horizons & New Frontiers

- IUPAC Identified 10 emerging chemical technologies in 2019
- Crystallography has a role to play in realising these, in particular:
  - Nanopesticides
  - Metal-Organic Frameworks for Water Harvesting (and a lot else)
  - Reactive extrusion (e.g. solid state reactions, via crystal grinding)
  - Solid State Batteries

The CSD contains lots of answers, and the right questions need to be asked of it.
Concluding thoughts and questions

- **Data growth**
  - CSD growth / structural chemistry data growth
  - Chemical space, diversity, coverage, trends over time
- **Advances** in instrumentation and techniques
  - Yielding more data with higher accuracy and precision
- **Automation** & Software development
  - Evolution of existing tool set / future tools development
- Where does the **community** best fit?
  - Publishing advances
- **Interdisciplinaritity** – connecting / linking other domains
  - Biology / Pharmacology / Medicine
- **Insights**
  - AI / ML based
  - From advanced search & visualisations
  - Derived from underlying high-quality data
- **New science & innovation**
  - Powered by scientific hypotheses underpinned by data
  - Spotting emerging trends
  - Digital Drug Design & Manufacturing Centres
- **Semantic** representation & analyses
  - Machine / human experts - symbiosis
  - Reasoning / Inferencing (based on ontologies)
- **Quality and precision of data and meta data**
  - Evolution of data standards, and data architecture (more capabilities & flexibility)
  - Data initiatives (FAIR, InChI, PIDs, CIF2.0)
  - Bulk operations: integrity checks, quality analyses, data metrics
Acknowledgements and thanks

Thank you and...

The CCDC Team(s)
Past and present members of the CCDC
The 396,937 authors of structures
All the many contributors to the CSD
And the whole scientific community and our various academic and industry partners
Olga Kennard for her original amazing vision

- Jason Cole
- Ian Bruno
- Suzanna Ward
- and others

for various slide contributions and helpful discussions

It’s a wonderful scientific journey – let’s keep walking on it, and celebrating key milestones and findings along the way!

http://www.ccdc.cam.ac.uk/
@ccdc_cambridge
ccdc.cambridge
The Cambridge Crystallographic Data Centre

- **International Data Repository**
  - Archive of crystal structure data
  - High quality scientific database

- **Scientific Software Provider**
  - Search/analysis/visualisation tools
  - Scientific applications

- **Collaborative Research Organisation**
  - New methodologies
  - Fundamental research

- **Education and Outreach**
  - Conferences, Workshops,
    Training, Teaching

**Enriching Chemistry with Crystallographic Data and Knowledge**

Dedicated to the advancement of structural chemistry and crystallography for the public benefit through providing high quality data, information services and software.

@ccdc_cambridge
ccdc.cambridge
http://www.ccdc.cam.ac.uk/

Thank you!