Database-driven discovery

Suzanna Ward and Eric Rogers

The Cambridge Crystallographic Data Centre
Creation of the CSD
Can Structural Knowledge Mitigate Risk?

Different interactions  Different solubility  Different stability

The Cambridge Structural Database (CSD)

- 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

An N-heterocycle produced by a chalcogen-bonding catalyst.

XOPCAJ – CSD one million
CSD One Million

A million thanks

The Cambridge Structural Database reaches one million structures, leading the way in structural data to inform drug discovery and materials development.

Cambridge, UK, 6 June, 2019. CSD (The Cambridge Structural Database) contains over 500,000 substances and over 1 million crystal structures. The database is one of the world's most complete resources on structural chemistry, providing a wealth of information about the structures of molecules and materials, and is used by researchers around the world.

The CSD is a unique resource for structural chemists and materials scientists, providing a wealth of information about the structures of molecules and materials. It is used by researchers around the world to help them understand how molecules behave and interact at the molecular level, and to design new materials with specific properties.

The CSD contains over 1 million structures, including small molecules, macromolecules, and materials. It is a valuable resource for researchers in a wide range of fields, including chemistry, materials science, and biology.

The CSD is a collection of high-quality data on the structures of molecules and materials, and is used by researchers around the world to help them understand how molecules behave and interact at the molecular level, and to design new materials with specific properties.

Thank you to the community

The CSD is a community-driven resource, and the growth of the database is due to the contributions of the users. We would like to take this opportunity to thank all of the users who have helped make the CSD a valuable resource.

Frequently asked questions

What does one million structures mean? It means that the CSD has a comprehensive collection of structures that can be used for research and development.

How can I access the CSD? The CSD is available online through a variety of search engines and databases, such as PubChem and the National Institutes of Health Enzyme Data Bank.

What is the future of the CSD? The CSD will continue to grow and evolve, with new structures being added regularly. We are committed to ensuring that the CSD remains a valuable resource for researchers around the world.
We clearly recognised even in those early days, that data banks have three principal functions. Firstly they must gather together existing knowledge and make it readily available to the scientific community. Secondly they can be used to reduce a large number of observations to a small set of constants and rules, and in this way transform a data base to a knowledge base. Such a knowledge base may obviate the need for further individual experiments in specific areas. Thirdly, they facilitate the comparison and collective analysis of individual results to gain insight into new or as yet unexplained phenomena. These ideas have been at the heart of the work of the Cambridge Crystallographic Data Centre and the driving force for improving methods of data collection, storage and dissemination. Most importantly they influenced development of computer programs and methodologies which are needed for the analysis and transformation of the accumulated information. (5)
Inside the CSD

Organic 43%

Metal-Organic 57%
At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Ti, Ge, Sn, Pb, Sb, Bi, Po

Not Polymeric 89%

Polymeric: 11%

Single Component 56%

Multi Component 44%

- 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
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 decisions.

Many of the most pressing challenges facing AI today revolve around its poor-quality training data. Bias, brittleness, ease of fooling, lack of representational edge case examples to fall back upon: all of these key problems trace their roots at least in part to poor quality training data. While algorithmic improvements could help, so too could having proper training data.

Curating the CSD

- Each dataset expertly curated
- Datasets enhanced
  - Chemical connectivity
  - Compound names
  - 2D chemical diagrams
  - Additional experimental data
  - Bibliographic information
Depositing the Data
Guidelines

The CCDC CIF Deposition Guidelines

When preparing your CIF for deposition please include as much information as possible and check it carefully. This is especially true for CSD Communications where there is no paper to describe the chemistry and experimental details leading to your structure. If you choose to publish your data as a CSD Communication please remember to provide all the authors/crystallographers/chemists who contributed to the crystallographic experiment as authors of the data. If we are unable to validate your structure from the information you have provided we may contact you. If we cannot resolve the issue, unfortunately, we may not be able to add your structure to the CSD.

Guidelines in Chinese

All experimental CIF files (including those from powder diffraction experiments) should contain an R-factor. This should be consistent with the crystallography being performed correctly and to the best ability that would be expected from the material and equipment used. We would like all experimental CIFs to contain:

- R-factors (R1, wR2, Rint)
- GooF
- Shift/ESD (to show that the refinement has converged)
- Explanation of any problems with numbers of reflections and parameters
- Any residual electron density
- Details of squeeze/solvent masking
- Atomic Displacement Parameter (ADP) values
- Temperature – cell and data collection temperatures match
- Experimental set up including mounting device and instrument type
- HKL included
- RES included

We would encourage you to take advantage of the IUCr checkCIF reports built in to the deposition page. This can highlight issues to check with your structure that can be clarified in the validation reply form, particularly in the case of A- or B-level alerts. Ideally, treatment of disorder or partial occupancy atoms should be clear and of course, no non-positive definite atoms!

To allow us to create the most accurate representation of your structure please provide as much additional information on the "Enhanced Data" page as is appropriate for your structure. Some chemical issues we commonly encounter when processing data into the CSD are:

- Given formula and crystal formula don’t agree. Particular attention should be paid to hydrogen atoms which may not be located in the experiment. It would be very helpful to us to have a complete moity formula (including unlocated hydrogens and any SQUEEZE/MASK species not located, if known)
- Charge balance, particularly for variable metal oxidation states and radicals
- Missing hydrogen atoms, especially on oxygen atoms that could be hydroxy/oxy/aqua ligands and for polyoxometalate structures
- Unusual bonding, tautomerics or metal-metal bonding
- Poorly handled or unmodelled disorder
- Unexplained void space not accounted for by SQUEEZE or MASK procedures

Further information that will benefit the users of your structure and that will enable the correct identification of any previous versions of your structure are:

- Stereochemical determination method, if relevant
- Crystalisation solvent/conditions
- Melting point
- Details of re-refinement – please tell us if the structure is a re-refined version of an existing CSD entry
- Rfactors or CCDC numbers of any known related structures, i.e. by temperature / stereochemistry / pressure, e.g. "high temperature determination of REPCODE"

If you have any further queries, please contact us via our Enquiries Page.
Adoption by the Community

% CIFs containing HKL data in CSD deposits

% Total CSD entries

Publication year


65%
What Else Could We Do?

• **Improved peer review**
  - Mandate crystallographic review of all structure-containing papers
  - Educate reviewers on nature of CheckCIF alerts

• **File requirements**
  - CIF + structure factors
  - Refinement instructions?
  - CheckCIF report?

• **Validation checks**
  - CheckCIF integration
  - Unit cell checks (with HKL checks? Or chemistry check?)
  - Geometry analysis?

• **Additional files available to reviewers?**
Over 180,000 entries from the Inorganic Crystal Structure Database (ICSD) now available through Access Structures.
Curation and Chemistry Assignment

Deposited CIF

CSD Entry
Using the CSD to Help With Curation

An automated probabilistic approach using data in the CSD

Assignment of chemistry is required to make data findable, interoperable and reusable

\[
P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}
\]

DOI: 10.1107/S0108768111024608
Challenges

- Missing atoms
- Element assignment
- Disorder
- Poor geometries
The Human Touch

- Each entry looked at by expert Scientific Editors
- Reliability scores focusses editorial efforts
- Manual validation of automated chemical interpretations improves automated methods

https://www.ccdc.cam.ac.uk/Community/blog/CSD-data-curation-the-human-touch/
Revisiting Data

Targeted improvements allow improved integrity, consistency, discoverability and value of data.

Ensure standardisation of early CSD entries

Creation and maintenance of subsets

Enrichment of data

Oxidation states
Melting Points in the CSD

>170,000 Melting Points

Study Temperature relative to MP
before additional CCDC validation

Study Temperature relative to MP
after additional CCDC validation
Maintaining Data Integrity

• **Integrity** – Completeness, consistency and trustworthiness

• **Data completeness** – Trends in reporting of metadata
  - Identify CSD Deposit checks and enhancements
  - Identify new filters to allow CSD users to better select fit for purpose data

• **Consistency** – Looking at experimental metadata to identify trends in information supplied

• **Trustworthiness** – Establishing automatic identification of potential cases of misconduct – including fraudulent and plagiarised data

Research integrity is much more than misconduct. *Nature*, 2019, 570, 5-5. DOI:10.1038/d41586-019-01727-0
Following Standard Ethical Practises

• CCDC is now a Member of the Committee on Publication Ethics.

• COPE’s objective is "to educate and advance knowledge in methods of safeguarding the integrity of the scholarly record for the benefit of the public".

• Membership gives us access to COPE resources and COPE advice – helping us deal with publication ethics and data integrity and issues.

https://publicationethics.org/about/governance
Making Crystallographic Data FAIR
CCDC database workflows
FAIR Data Principles

Findable, Accessible, Interoperable, Reusable
FAIR Data Principles

The FAIR Guiding Principles for scientific data management and stewardship

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


Findable

Accessible

Interoperable

Reusable

“all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people” (Wilkinson, M. D. et al., 2016: 3)

http://www.datafairport.org/
Plan S

“Although the Plan S principles refer to peer-reviewed scholarly publications, cOAlition S also strongly encourages that research data and other research outputs are made as open as possible and as closed as necessary.”

https://www.coalition-s.org/principles-and-implementation/

European Commission

“...the implementation of FAIR data needs to go hand-in-hand with the principle that data created by publicly-funded research must be as Open as possible and as closed as necessary. The EC and Member States should consider FAIR and Open as complementary concepts and address both in policy.” (EU Commission, 2018 : 10)

“Since 2017, all Horizon 2020 projects are part of the Open Research Data Pilot by default. The Principal Investigator must:

• Develop a data management plan in the first 6 months of the project and keep it up-to-date throughout their project;
• Deposit their research data in a suitable research data repository;
• Make sure third parties can freely access, mine, exploit, reproduce and disseminate their data;
• Make clear what tools will be needed to use the raw data to validate research results, or provide the tools themselves.”

“ERC beneficiaries are encouraged to take part in the H2020 Open Research Data Pilot, but this is not compulsory. Those who take part in the Open Research Data Pilot must adhere to the obligations outlined above.”

https://www.data.cam.ac.uk/funders
Publisher research data policies

“The Royal Society of Chemistry believes that, where possible, all data associated with the research in a manuscript should be freely available in an accessible and usable format, enabling other researchers to replicate and build on that research. Therefore, in addition to providing the data required for submission (as detailed above) we encourage authors to deposit as much data as possible that is related to the research in their article. This should be in appropriate and publicly available repositories.”

https://www.rsc.org/journals-books-databases/journal-authors-reviewers/prepare-your-article/experimental-data/

“It is the practice of IUCr journals to provide free access to all supplementary materials and supporting data files deposited with a published article.”

https://journals.iucr.org/services/authorrights.html
Aspects of FAIR Data

**Findable**
- Globally unique and persistent identifiers
- Rich metadata descriptions
- (Meta)data available in a searchable resource

**Interoperable**
- Standard formats for representation
- Use of FAIR vocabularies
- References to other (meta)data

**Accessible**
- (Meta)data retrievable by their identifier
- Standard, open communication protocols
- Metadata accessible even when data are not

**Reusable**
- Described with a plurality of attributes
  - data usage licenses
  - detailed provenance
  - domain-relevant community standards
Data Repositories

Role of Repositories in making data FAIR:
- Long-term data preservation
- Continued access to data
- Assignment of identifiers and DOIs
- Added descriptive metadata
- Searchable databases
- Data curation

Repository frameworks and systems

- DOI assignment
- Open Archival Information System (OAIS) Reference Model
- Trusted Repository Certification
- Global and domain specific membership groups
- Data preservation policies & plans

Data Repositories

How to find a repository for your research data

- Search a repository registry
  - https://www.re3data.org/

- Use a repository recommended by the publisher or funding body
- Search for repositories among accreditation bodies
  - https://www.coretrustseal.org/
  - https://www.icsu-wds.org/services/certification
Crystallographic Information File: CIF

A standard format for archive and exchange of crystallographic data

- derived model
- processed data (structure factors)
- metadata about raw data (imgCIF)

International Union of Crystallography
Commission on Crystallographic Data
Commission on Journals
Working Party on Crystallographic Information

The Crystallographic Information File (CIF): a New Standard Archive File for Crystallographers*

BY SYDNEY R. HALL
Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

FRANK H. ALLEN
Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND I. DAVID BROWN
Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada

(Received 9 April 1995, accepted 26 June 1995)
CIF as a FAIR data format

**Findable**
- Searchable fields for identifiers and metadata descriptions

**Interoperable**
- Standard dictionary and vocabularies
- Standard format for processed and derived data

**Accessible**
- Trusted searchable data repositories:
  - Cambridge Structural Database
  - Inorganic Crystal Structure Database
  - Protein Data Bank

**Reusable**
- Data provenance
- Software packages and parameters
- Quality metrics
CCDC Dataset Workflow

Deposition (Pre-ingest) → Deposition (Ingest) → Publication → Data Access → Curation
Data Deposition (Pre-ingest)

- Manual User Actions:
  - Provide personal details and upload files
  - Fix syntax errors (if any)
  - Explain why no structure factor data
  - Add explanations for checkCIF alerts
  - Add crystallographer details
  - Provide known publication details
  - Add additional scientific metadata
  - Review and confirm

- Automated Actions:
  - Syntax check
  - Generation of checkCIF report
Data Deposition (Ingest)

- **Automated Actions:**
  - Syntax check
  - Check for duplicate dataset
  - Internal record creation
  - Assigning identifiers

- **Manual CCDC Actions:**
  - Dealing with non-standard file formats
  - Investigating duplicate datasets
  - Updating records with resubmitted data
Data Publication

- Pre-publication metadata communicated to CCDC by journal publishers
- Full publication metadata communicated to CCDC by journal publishers
- Publication information updated via journal scanning by CCDC staff

Pre-publication metadata added to data record

Full publication metadata added to data record

CCDC checks run on deposited files:
- Pre-ingest
- Full publication metadata added to data record

Publication information updated via journal scanning by CCDC staff

Full publication metadata communicated to CCDC by journal publishers

Tabernabovines A–C: Three Monoterpenoid Indole Alkaloids from the Leaves of Tabernaemontana bovina

Yang Yu, Mei-Fen Bao, Jing Wu, Jing Chen, Yu-Rong Yang, Johann Schinnerl, and Xiang-Hai Cai

Accession Codes

CCDC 1916676 contains the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.
Data Publication

**Sources of Publication Information:**

- Pre-publication metadata communicated by journal publisher feeds
- Full publication metadata communicated and updated by journal publisher feeds

**Manual CCDC Actions:**

- Reviewing publication details
- Publication information updated via journal scanning by CCDC staff
- Publication information communicated by researchers wanting to access data
Data Publication

Accession Codes

CCDC 1916676 contains the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.
Data Access

- Data is made accessible:
  - Pre-publication to reviewers and depositors to facilitate preparation of manuscripts
  - Immediately through Access Structures once data is published
  - Through the CSD once curated by CCDC’s scientific editors.

- Processes for making data more findable and interoperable:
  - Identifiers added to data entries
  - Links from publication articles to data created
  - Links from CCDC datasets to other databases
Standard Identifiers and Interoperability

Data should be considered legitimate, citable products of research...

Dataset Publication
http://dx.doi.org/10.5517/ccngvdb

- The CCDC registers DOIs for datasets through DataCite
- Metadata for CCDC datasets is openly accessible via DataCite
- Foundation for interoperability and formalising data citation

ORCID IDs for Researchers
At least 30% of current CSD depositors provide an ORCID ID

Andrew Bond
ORCID ID
https://orcid.org/0000-0002-1744-0486
Links from Articles to CCDC Data

This journal is © The Royal Society of Chemistry 2009

Wiley Online Library

An efficient phosphate sensor: trippodal quinoline excimer transduction

Elsevier

SCOLIX

Research data for this article

Cambridge Crystallographic Data Centre

Crystallographic data

Data associated with the article:
CCDC 1543805: Experimental Crystal Structure Determination

Outlines

Abstract
Graphical abstract
1. Introduction
2. Results and discussion
3. Conclusions
4. Experimental
Acknowledgements
Supplementary data
References and notes

X CONTENT

CCDC 1543805 (1543805) contains the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre.
Making data accessible through the CSD

Data not published in a scientific journal can be curated into the CSD and made available to the community as a CSD Communication.

Structures from your PhD thesis can be made publicly available through the CSD.

https://www.ccdc.cam.ac.uk/Community/csd-communications/
Data Curation

- A reliable chemical representation is essential for enabling reuse and application of crystallographic data
- Representation is generated at CCDC using a combination of automated processes and manual validation

Assignment of chemistry is required to make data findable, interoperable and reusable
Enablers of FAIR Crystallographic Data

- **Standard formats and identifiers**
  - Crystallographic Information Format and dictionaries (CIF)
  - Standard Identifiers and associated infrastructure (DOIs, ORCID, InChI…)

- **Community stakeholders**
  - Instrument providers and software developers adopting standards
  - Publishers and editors encouraging use of standards for publication
  - Repositories and databases providing access to enriched data
  - International Unions supporting and promoting standards
  - Individual researchers and others championing research data standards

- **Tools and services that make it easy to make data FAIR**
From Data to Knowledge
From Experiment to Knowledge

Experiment

C₁₀H₁₇N⁺, Cl⁻

CCDC

Knowledge

Rhinovir: An Extraordinary Example of Conformational Polymorphism

Abstract

Proposed: In the manner of solid-state molecular crystal syntheses that demand exact structures, and the solution of a new crystal has valid the replication of the same. The paper presents characteristics of the new polymorph and the structure and hydrogen bonding network for each layer.
Advanced Services and Software

CSD-Enterprise
All CCDC applications and software

CSD-Discovery
To discover new molecules with pharmaceutical applications

CSD-System
To search, visualise, analyse and communicate structural data

CSD-Materials
To understand and predict solid form stability and properties

The Cambridge Structural Database
Generating Insights

• The **CSD Python API** enables you to create tailored scripts using full array of CSD functionality

• Answer targeted research questions or integrate access with other software

• Functions include:
  • Full search capabilities
  • Geometry analysis
  • Interaction analysis
  • Descriptor calculation
  • 2D diagram generation
Increasing Complexity

- Increasing:
  - Formula weights
  - Unit Cells
  - Number of elements
Trends in Experimentation

The graph illustrates the percentage of various technologies used in experimentation over the years from 2000 to 2018. The technologies include:

- point%
- image plate%
- area + multiwire%
- pad%
- cmos%

The graph shows a general decline in the use of point% technology from 100% in 2000 to approximately 0% in 2018, with a corresponding increase in the use of other technologies.
Elements in the CSD

Percentage of structures that contain each element from before (red) and after (yellow) 2009
Drugs

- Top 200 Pharmaceutical Products
  - By retail sales in 2018
  - Produced by the Njarðarson Group
  - The University of Arizona
  - Drugs already in the CSD coloured green

- *J. Chem.* Ed. 2010, 87, 1348
Identifying Trends in Drug Structures

https://www.ccdc.cam.ac.uk/Community/blog/insights-into-drug-like-compounds-from-crystal-data/
The CSD and the PDB

Linking
- Between CSD and PDB ligands

CSD-CrossMiner
- Pharmacophore query tool
- Searches the CSD and PDB
Using the Data

Match PDB ligands to best representative CSD molecules

Date correspondences between the PDB and CSD archives now available

The Wiley Protein Data Bank and the Cambridge Crystallographic Data Centre (CSD) [http://www.ccdc.cam.ac.uk] are pleased to announce the availability of a new database containing correspondence between the PDB structure components and ligand molecules found in the PDB archive that exactly match small molecules in ray structures in the Cambridge Structural Database (CSD) archive.

The chemical structure of every unique molecule in the Protein Data Bank is described in the PDB Chemical Component Dictionary. The new PDB Chemical Component listing data the complements information in the PDB by providing the following CSD information for matching molecular entities: accession code, descriptors, database, compound name, and resolution. The PDB Chemical Component Dictionary is available for the PDB database via the "Chemical Component Dictionary" tab in the "Secondary Structure" section. The data is available for download in a tab-separated text file format.

As of the latest update, there are 20,677 chemical components listed in the PDB Chemical Component Dictionary, and over 1,110 of these PDB chemical components have been identified in the PDB. The new PDB Chemical Component Dictionary can be accessed via the "Chemical Component Dictionary" tab in the "Secondary Structure" section.
Solid Form Informatics

- The term “solid form informatics” first introduced in mid-2000s by Bob Docherty (Pfizer):
  Use of structural knowledge to inform key decisions in pharmaceutical development
- Solid form informatics now a key part of the solid form development workflow at most major pharmaceutical companies
From Data to Knowledge

Individual data points from different datasets combine to provide information that aids in the discovery and optimisation of new chemical entities.

Predicting Unlikely Interactions

Predictive analytics is used to identify the likelihood of specific molecular interactions occurring from similar crystal structures.

The integration of solid-form informatics into solid-form selection

One in half a million: a solid form informatics study of a pharmaceutical crystal structure
Peter T. A. Galek*, Elna Pickock, Peter A. Wood, Ian J. Bruno and Colin R. Groom*

Navigating the Solid Form Landscape with Structural Informatics
Peter T. A. Galek, Elna Pickock, Peter A. Wood, Neil Feeder, Frank H. Allen
Book Editor(s): Yuriy A. Abramov

Knowledge-based H-bond prediction to aid experimental polymorph screening
Peter T. A. Galek, Frank H. Allen, László Fabián and Neil Feeder

Galek et al, CrystEngComm, (2009), 11, 2634 - 2639
CSD-Materials: Targeted Solutions

- Crystal Packing Similarity
- Motif Search & Packing Feature Search
- Conformer Generator
- Full Interaction Maps
- DASH
- Complex Structural Analysis
- Solid Form Risk Assessment
- Solid Form Design
- Hydrate Analyser & Solvate Analyser
- Calculations
- Hydrogen Bond Propensity
- Molecular Complementarity
“A search of the Cambridge Structural Database using a series of pharmacophore queries led to the discovery of an O-spiroketal C-arylglucoside scaffold. Subsequent chemical examination combined with computational modelling resulted in the identification of the clinical candidate 16d (CSG452, tofogliflozin), which is currently under phase III clinical trials.”

Yoshihito Ohtake et al Journal of Medicinal Chemistry 2012 55 (17), 7828-7840 (Roche, Chugai)
Collective value

Articles

The Supramolecular Synthon Approach to Crystal Structure Prediction

J. A. R. P. Sarma* and Gautam R. Desiraju**

gsk bioSciences Pvt. Ltd., #210, ‘My Home Tycoon’, 6-3-1192, Begumpet, Hyderabad 500 016, India, and School of Chemistry, University of Hyderabad, Hyderabad 500 046, India

Received December 21, 2001: Revised Manuscript Received January 11, 2002

© This paper contains enhanced objects available on the Internet at http://pubs.acs.org/crystal.

ABSTRACT: A new approach has been proposed for the ab initio crystal structure prediction of small organic molecules. This exercise forms a part of the recent blind test on crystal structure prediction conducted by the Cambridge Crystallographic Data Centre. The method uses a starting point lists of low energy structures generated by an exhaustive computational procedure, namely, the Polymorph Predictor program in Cerius². Such computational procedures take into account only the enthalpic factors in crystallization. A further difficulty is that information relating to crystallization kinetics is very hard to obtain directly. However, such kinetic information is implicitly contained in the experimental structures that are found in crystallographic databases. Therefore, in our approach, the low energy structures obtained in the Polymorph Predictor program are reranked after consideration of experimental structures of structurally similar molecules. Operationally, this is most conveniently carried out after identification of possible supramolecular synthons in the Cambridge Structural Database. These synthons are representative structural units that convey critical information that relates isolated molecules with their resulting crystal structures. Of the three molecules in the blind test, the present approach was fully successful for one, but only of limited utility in the two others. Reasons for this variability of success are given.
Collective value

we note that if the CSD were to be significantly larger than what it is today, say, around a million refcodes, CSP with the synthon-based approach could be successfully employed for a much wider variety of molecules.

In summary, and from the viewpoint of CSP, the utilization of structural information could provide a more effective sieve toward the correct solution. As the amount of structural information in crystallographic databases increases, structure prediction would gradually move toward fingerprinting.
Using the Collection

CCDC Blind Test Showcases Major Advance in Crystal Structure Prediction Methods

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 of polymorphs, salts and hydrates. Experimental structures were predicted to be highly accurate.

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

November 2015

The structure of an organic molecule on a napkin may not be apparent that there are is of possible ways that it could assemble as a 3D crystal. Now, a collaboration of dozens of scientists and computer programmers has successfully predicted the crystal structure of five, ex, 'drug-like' organic molecules — using nothing but a 2D map showing which atoms fit to which.

The achievement, announced 27 October at a workshop in Cambridge, UK, paves the way for developments that would cut the cost of the design and manufacture of drugs and other chemicals, as well as further our understanding of fundamental chemistry.
Thank you