CODATA

Sustainability of life and molecular crystallography 3D data

John R. Helliwell

School of Chemistry, University of Manchester, UK

john.helliwell@manchester.ac.uk

Brian McMahon

IUCr, 5 Abbey Square, Chester CH1 2HU, UK

bm@iucr.org

SciDataCon2014: New Delhi, India, 4 November 2014

2014

international year of crystallography

United Nations Educational, Scientific and Cultural Organization · International Union of Crystallography

Partners for the International Year of Crystallography 2014

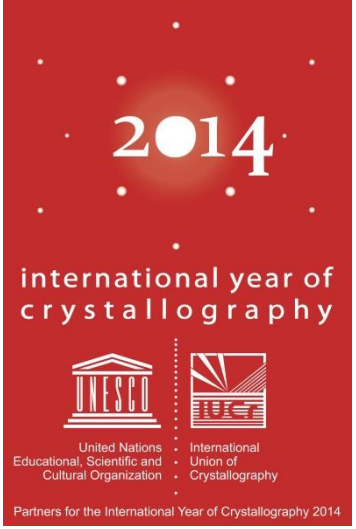
IYCr2014

UN General Assembly, GA/11262, 3 July 2012
Resolution 66/284
submitted by Morocco, approved unanimously

United Nations

2014 is proclaimed International Year of Crystallography

- ***Recognizing that humankind's understanding of the material nature of our world is grounded, in particular, in our knowledge of crystallography;***
- *Stressing that education about and the application of crystallography are **critical in addressing challenges such as diseases and environmental problems, as well as solutions for plant and soil contamination;***
- *Considering that the **impact** of crystallography is present everywhere in our daily lives;*
- *Considering also the significance of the **scientific achievements** of crystallography, as illustrated by twenty-three Nobel Prizes awarded in the area, and that **crystallography is still fertile ground** for new and promising fundamental research;*
- *Considering further that 2014 marks the centenary of the beginning of modern crystallography and its identification as the **most powerful tool** for structure determination of matter,*
- *Being aware that 2014 provides an opportunity to promote **international collaboration** as part of the sixty-fifth anniversary of the founding of the International Union of Crystallography;*
- *Noting the broader welcome by the crystallographic community worldwide of the idea of 2014 being designated as the International Year of Crystallography*



Crystallography in modern life

2014 is proclaimed International Year of Crystallography

United Nations

- *"Crystallography also has an important place as we work for inclusive sustainable development – policies that are good for people and the planet"*

  Ban Ki-Moon, UN Secretary-General • IYCr 2014 Opening Ceremony
- Crystallography has an important role within the topic of the sustainability of Life.
- It is an essential analytic tool in academia and industry.
- In health sciences, structure-based drug design is now employed.
- In energy research hydrogen storage is addressed also using 3D atomic structures.

Examples of crystallography in the life sciences

Based on IUCr website 11 September 2014

The Nobel Prize in Chemistry 1964
Dorothy Crowfoot Hodgkin

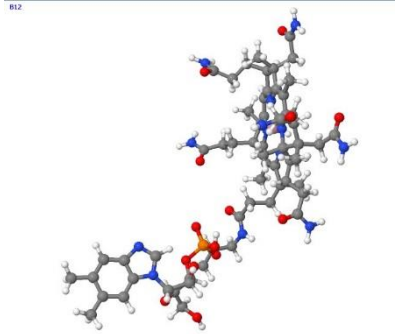
The Nobel Prize in Chemistry 1964

Dorothy Crowfoot Hodgkin

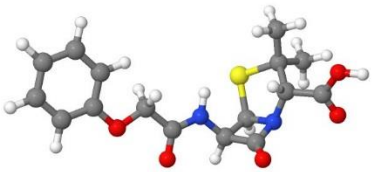
Prize share: 1/1

The Nobel Prize in Chemistry 1964 was awarded to Dorothy Crowfoot Hodgkin *"for her determinations by X-ray techniques of the structures of important biochemical substances"*.

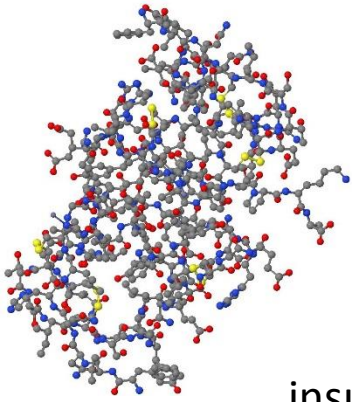
Dorothy Hodgkin (1910-1994): still a major inspiration today

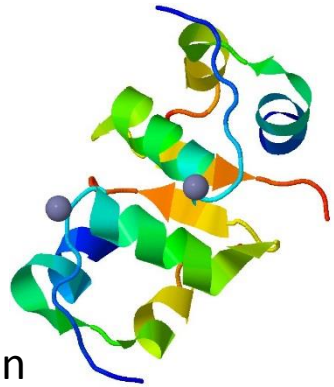


vitamin B12 (cobalamin)



penicillin

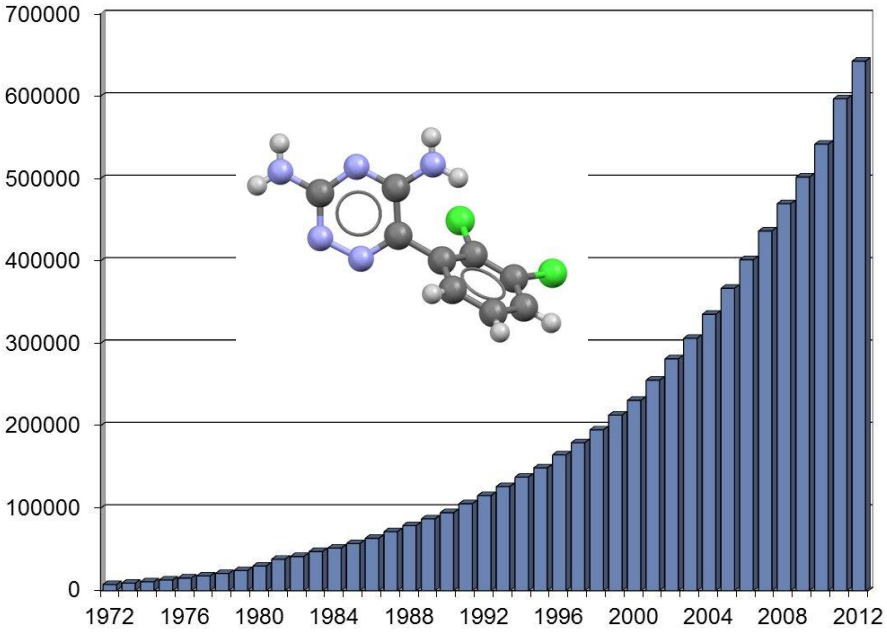




insulin

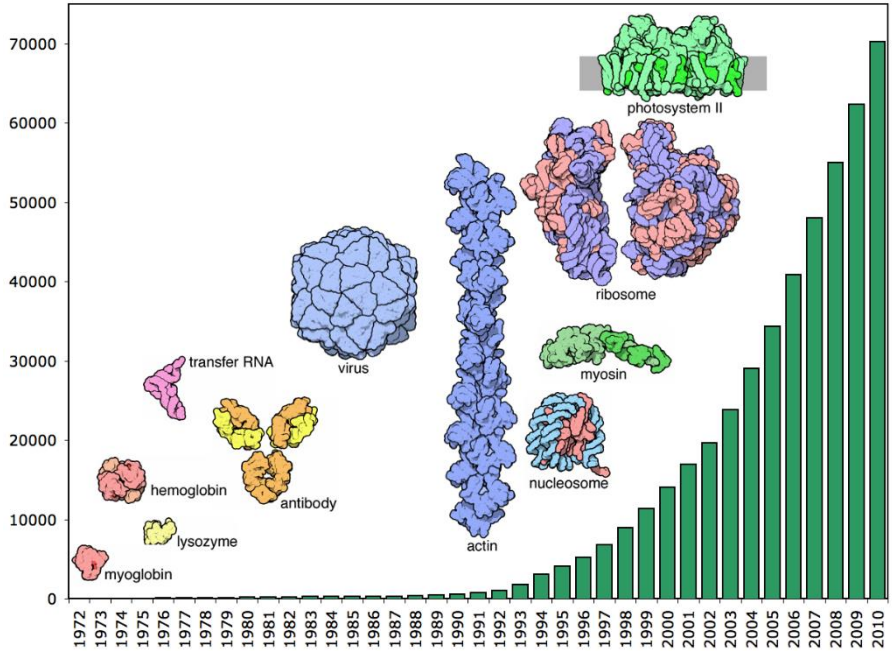
"The Nobel Prize in Chemistry 1964". *Nobelprize.org.* Nobel Media AB 2014. Web. 22 Oct 2014. <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1964/>

Structures large and small



Cambridge Structural Database

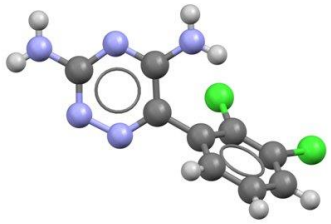
686944 structures at 6 January 2014



Protein Data Bank

104371 structures at 23 October 2014

Interactions between large and small

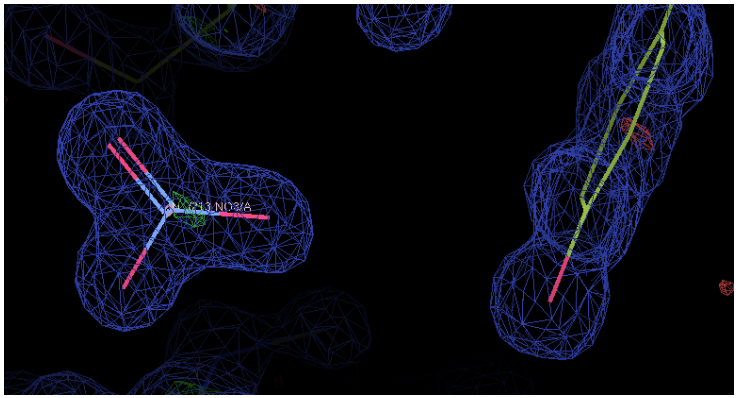


Lamotrigine
(500,000th structure in CCDC)

- Anti-epileptic drug
- Acts at voltage-sensitive sodium channels to stabilize neuronal membranes and inhibit the release of excitatory amino acid neurotransmitters
- **Relatively little binding to plasma proteins – relatively low toxicity**
- Binding to melanin may have long-term ophthalmologic implications

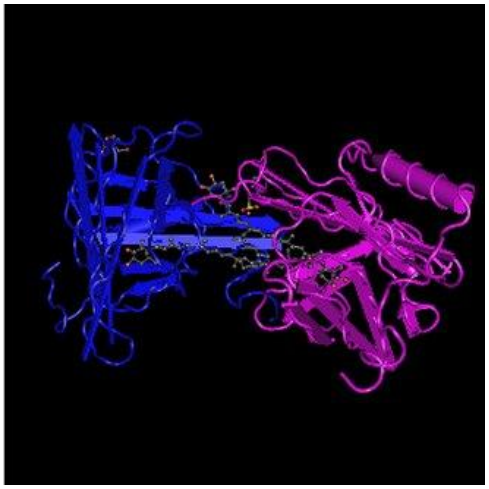
Chemicals in the environment

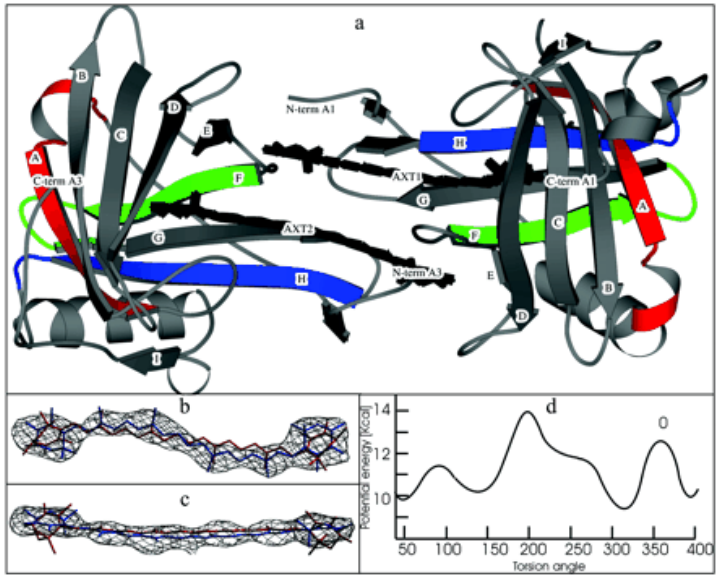
- Life under 'normal' and extreme conditions can be reliably compared; hot springs, high saline and extreme cold examples of protein crystal structures have been determined.
- It is a natural next step to understand the effects of pollution.



Discrete nitrate binding to a protein surface

Chemicals in the environment





Not exactly ***life*** at high temperatures…!

Cianci, M., Rizkallah, P., Olczak, A., Raftery, J., Chayen, N., Zagalsky, P. & Helliwell, J. (2002). The molecular basis of the coloration mechanism in lobster shell: β-crustacyanin at 3.2-Å resolution. *Proc. Natl Acad. Sci. USA*, **99,** 9795-9800

Life on Earth exists in a fairly restricted temperature range!

More, bigger, better…

The European Synchrotron Radiation Facility (ESRF) in Grenoble, in which the UK has a 14% share, and the Institut Laue Langevin nuclear reactor is to the right (UK share 25-33%)

ESRF: premier **X-ray** source

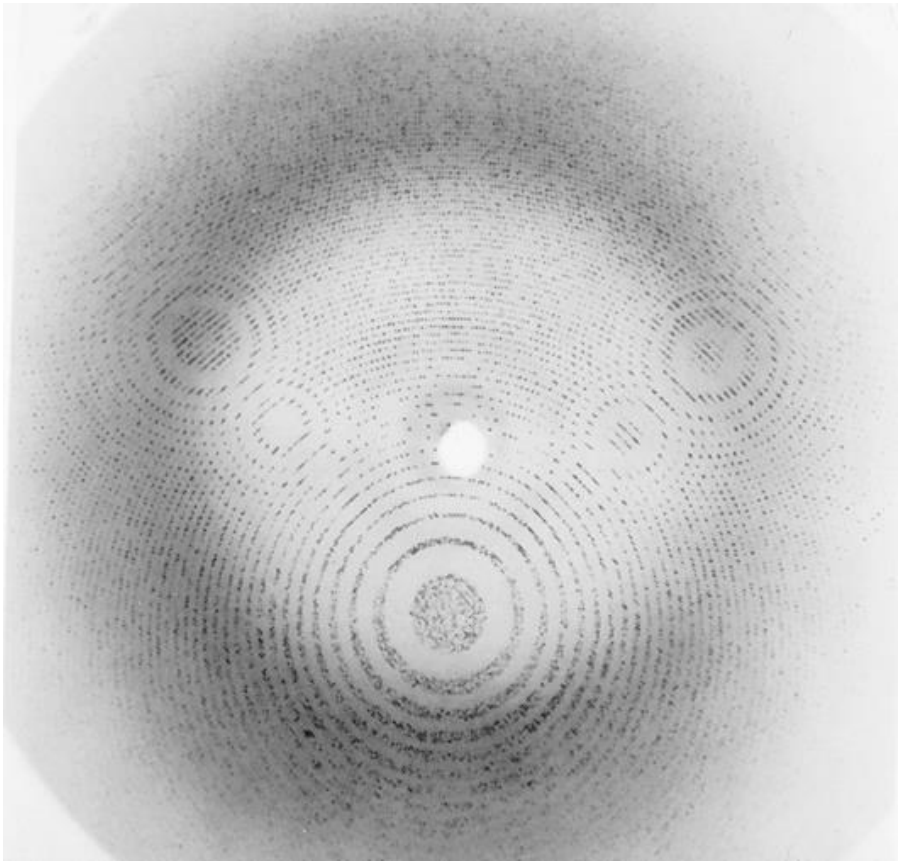
World leading

Institut Laue Langevin: **Neutron** source

John Helliwell has Chaired the ESRF Science Advisory Committee (SAC) and the Biological Crystallography Neutron Beam Committee; He currently Chairs the Spanish Synchrotron SAC.

… delivers more, better results

Rhinovirus oscillation photographic film exposure recorded at the Cornell SR source (courtesy Prof. M. G. Rossmann Purdue U., USA). These congested virus crystal patterns posed special challenges for detector developers.



NATURE VOL. 317 12 SEPTEMBER 1985 ARTICLES 145

Structure of a human common cold virus and functional relationship to other picornaviruses

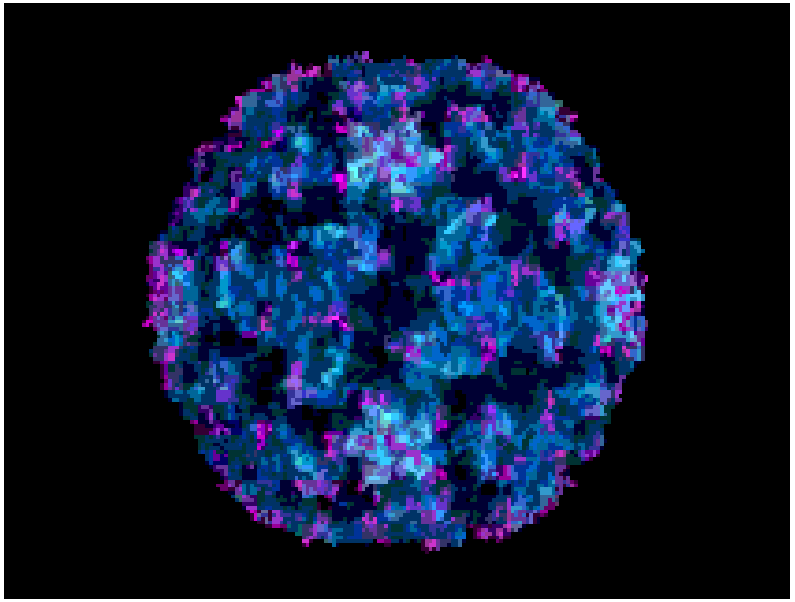
Michael G. Rossmann*, Edward Arnold*, John W. Erickson*†, Elizabeth A. Frankenberger*†, James P. Griffith*, Hans-Jürgen Hecht*†, John E. Johnson*, Greg Kamer*, Ming Luo*, Anne G. Mosser‡, Roland R. Rueckert‡, Barbara Sherry‡ & Gerrit Vriend*

* Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA
‡ Biophysics Lab, University of Wisconsin, 1525 Linden Drive, Madison, Wisconsin 53706, USA

Professor M. G. Rossmann, USA

In 1985, he became the first scientist to build a model of human rhinovirus-14, HRV-14, one of about 100 known cold virus strains. This was amongst the first group of virus crystal structures solved, and opened up studies of the most technically challenging projects.

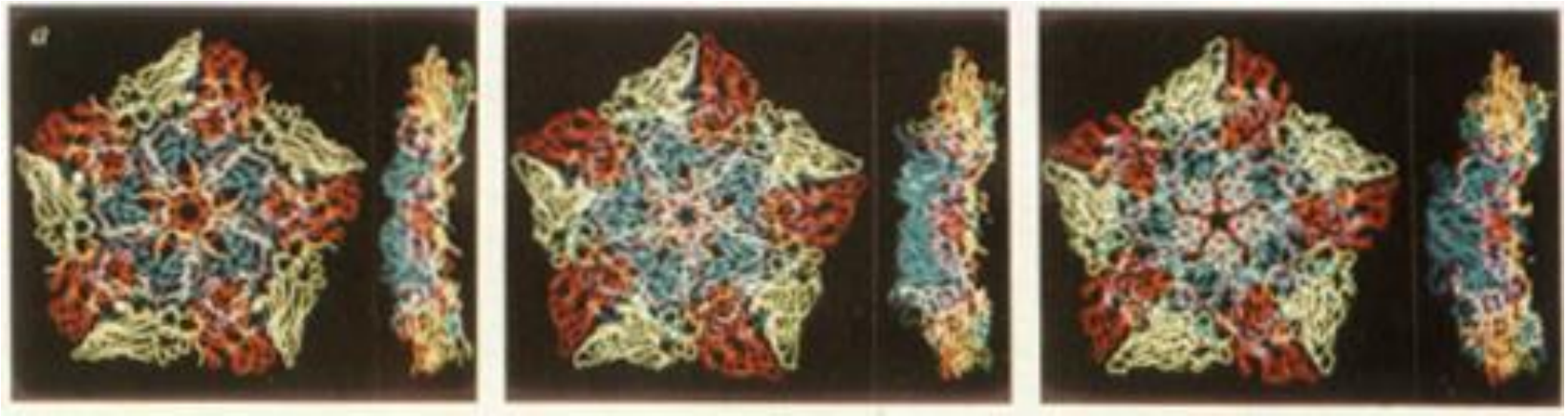
Virus 3-D structures



Human rhinovirus 14
Michael Rossmann, USA

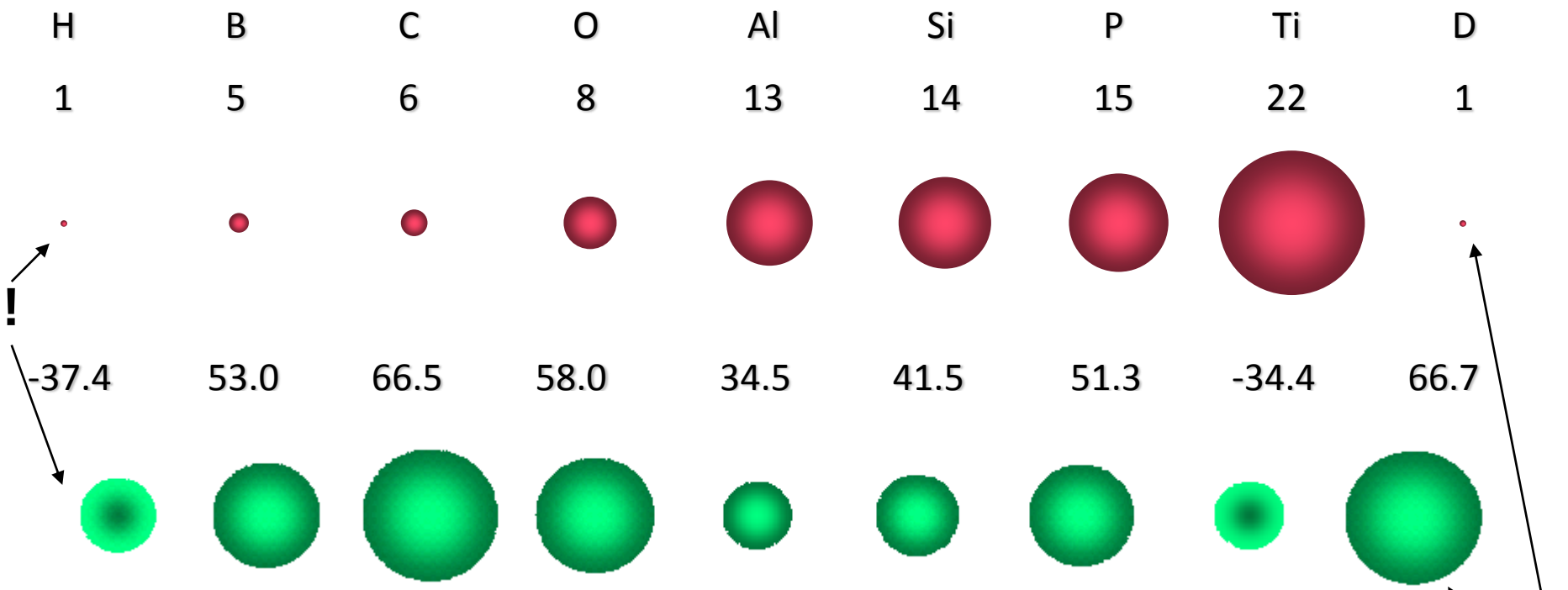
Comparing virus surfaces

FMDV | Mengo Virus | HRV 14



X-rays and neutrons

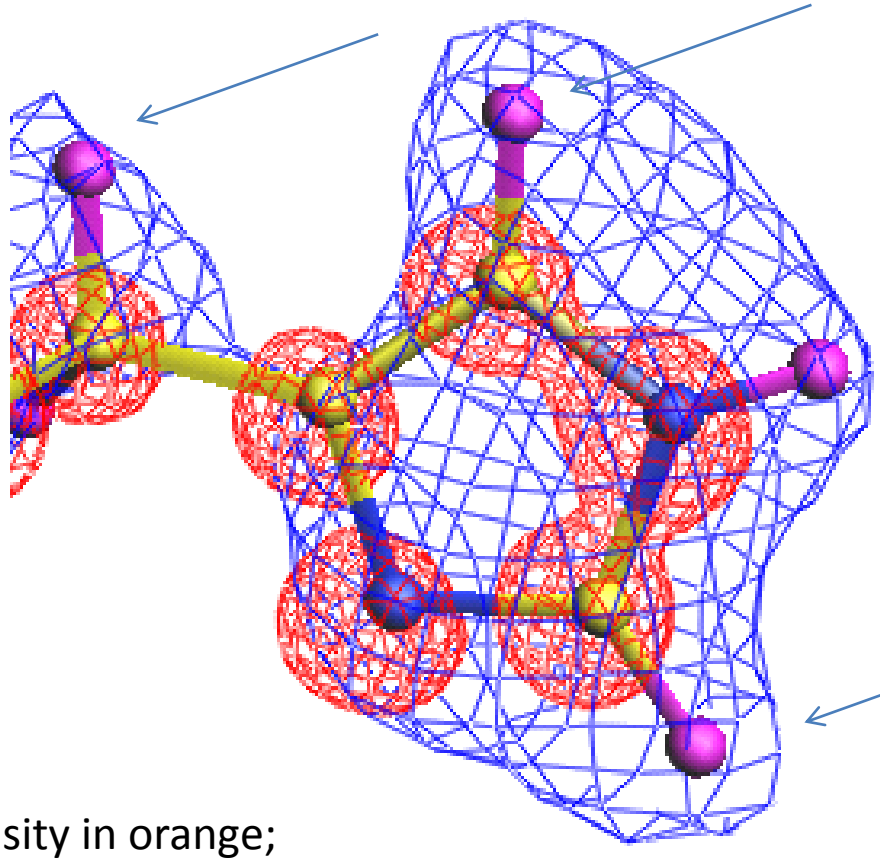
X-rays; Scattered from electrons proportional to Z (red). Neutrons; scattered from nuclei & evenly across all elements (green)



- **Large difference in the neutron cross-section among isotopes**
- **Neutron diffraction can be used to directly determine the positions of H-isotopes at medium resolutions (~2.5 Å)**

Neutron and X-ray crystallography of proteins

Slide prepared by Matthew. P. Blakeley, Institut Laue Langevin, Grenoble



As an example
a histidine amino
acid in a protein

X-ray electron density in orange;
Neutron nuclear density in blue (note the deuteriums are now visible)

What has this all to do with “data”?

- Structural models, stored in curated databases, have immense value for comparative studies, new compound discovery etc.
- Experimental data, stored as reduced and processed data sets, are invaluable for validating models and re-refinement of structures
- Raw experimental data has the potential for unleashing new methods and new science

The Nobel Prize in Physics 1915
William Bragg, Lawrence Bragg

The Nobel Prize in Physics 1915

Sir William Henry Bragg

Prize share: 1/2

William Lawrence Bragg

Prize share: 1/2

The Nobel Prize in Physics 1915 was awarded jointly to Sir William Henry Bragg and William Lawrence Bragg *"for their services in the analysis of crystal structure by means of X-rays"*

Lastly, though the absorption coefficient of the tungsten peak has not yet been satisfactorily measured, it may be doubtless supposed to be a little less than that of the A peak of platinum, since its wave-length is slightly less. A recent measurement of the latter quantity gives the value 35·5 and the absorption coefficient of the characteristic radiation of tungsten is given by Barkla as 33.

The Structure of Some Crystals as Indicated by their Diffraction of X-rays.

By W. L. Bragg, B.A.

(Communicated by Prof. W. H. Bragg, F.R.S. Received June 21,—Read June 26, 1913.)

[Plate 10.]

A new method of investigating the structure of a crystal has been afforded by the work of Laue* and his collaborators on the diffraction of X-rays by crystals. The phenomena which they were the first to investigate, and which have since been observed by many others, lend themselves readily to the explanation proposed by Laue, who supposed that electromagnetic waves of very short wave-lengths were diffracted by a set of small obstacles arranged on a regular point system in space. In analysing the interference pattern obtained with a zincblende crystal, Laue, in his original memoir, came to the conclusion that the primary radiation possessed a spectrum consisting of narrow bands, in fact, that it was composed of a series of six or seven approximately homogeneous wave trains.

In a recent paper† I tried to show that the need for assuming this complexity was avoided by the adoption of a point system for the cubic crystal of zincblende which differed from the system considered by Laue. I supposed the diffracting centres to be arranged in a simple cubic space lattice, the element of the pattern being a cube with a point at each corner, and one at the centre of each cube face. A simpler conception of the radiation then became possible. It might be looked on as continuous over a wide range of wave-lengths, or as a series of independent pulses, and there was no longer any need to assume the existence of lines or narrow bands in its spectrum.

* W. Friedrich, P. Knipping, and M. Laue, 'Münch. Ber.,' June, 1912.
† 'Camb. Phil. Soc. Proc.,' November, 1912.

W. L. Bragg. The Structure of Some Crystals as Indicated by Their Diffraction of X-rays. *Proc. R. Soc A,* 1913 **89**: 248-277.

"The Nobel Prize in Physics 1915". *Nobelprize.org.* Nobel Media AB 2014. Web. 22 Oct 2014. <http://www.nobelprize.org/nobel_prizes/physics/laureates/1915/>

Key terminology – raw data

- 1913

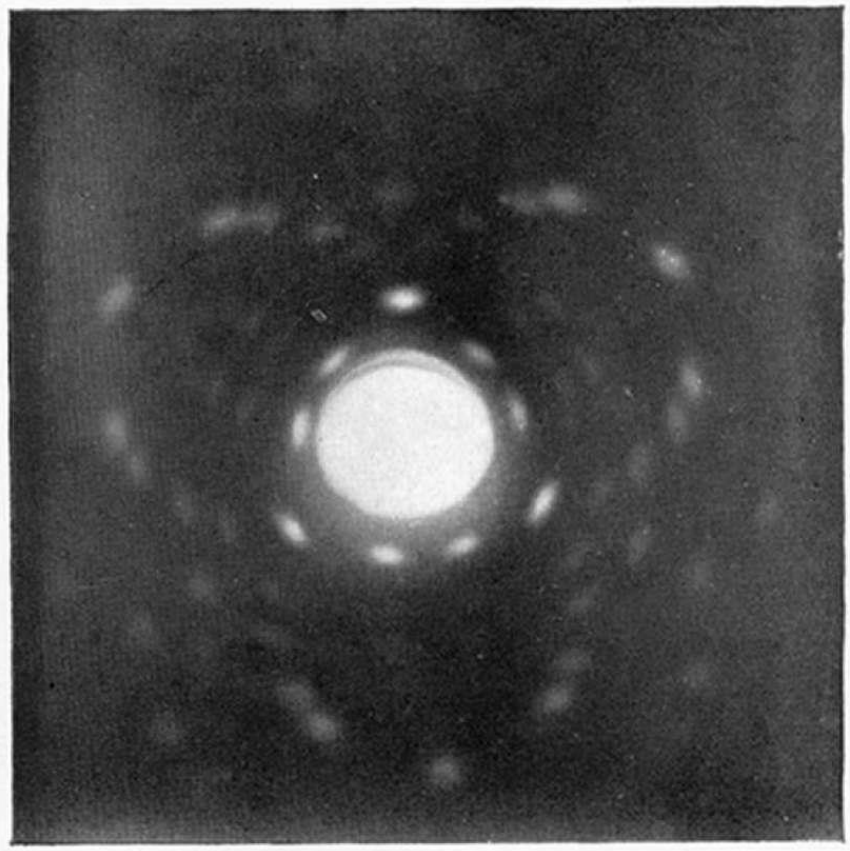
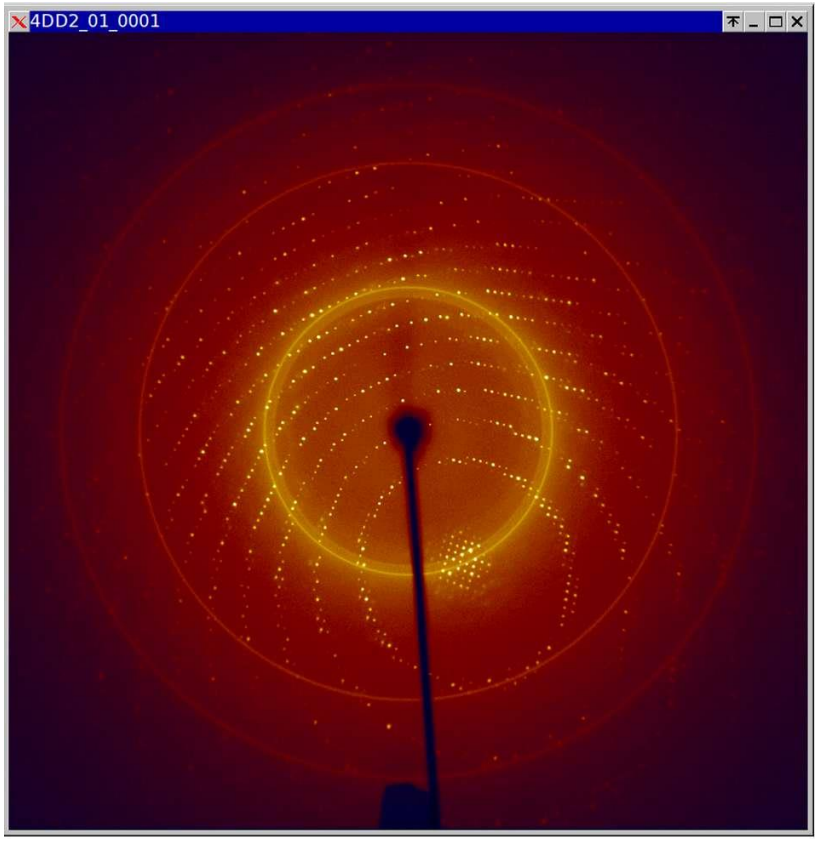


Fig. 11.—Fluorspar.

Bragg, W. L. (1913). The Structure of Some Crystals as Indicated by their Diffraction of X-rays. *Proc. R. Soc. London Ser. A*, **89**, 248-277.

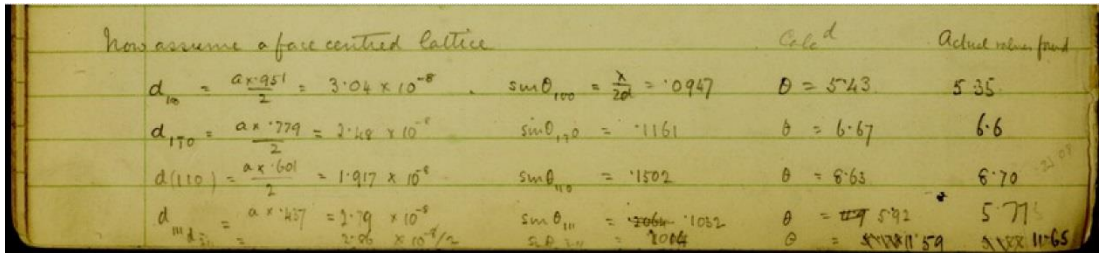
- 2013



Tanley, S. W. M., Schreurs, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals. *J. Appl. Cryst.* **46**, 108-119.

Key terminology – processed data

- 1913



Plane.	Calculated Spacing.	Calculated Angle.	Observed Angle.
100	$d_{(100)} = 3.04.$	5° 26′	5° 21′
$1\bar{1}0$	$d_{(1\bar{1}0)} = 2.48.$	6° 40′	6° 36′
110	$d_{(110)} = 1.92.$	8° 38′	8° 42′
111	$d_{(111)} = 2.79.$	5° 55′	5° 46′
$2\bar{1}\bar{1}$	$d_{(2\bar{1}\bar{1})} = 1.43.$	11° 35′	11° 39′

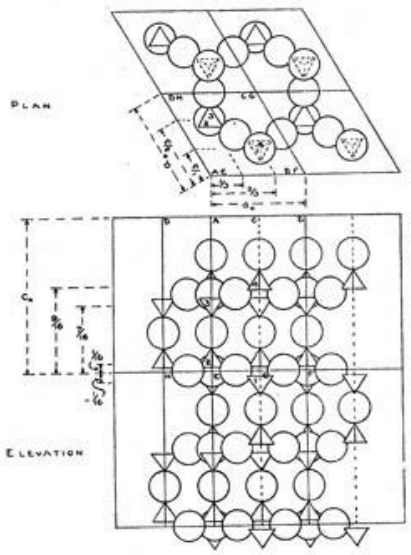
The upper figure is from the Braggs' notebook, showing observed angular locations of diffracted beams from different crystal planes, and their calculate values. The lower figure is a similar published tabulation from the 1915 book *X-rays and Crystal Structure*.

- 2002

```
# h,k,l, Fo-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_reflns_list_code          4
_shelx_F_calc_maximum       183.83
_exptl_crystal_F_000       1144.00
_reflns_d_resolution_high  0.7705

loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, y+1/2, -z+1/2'
 '-x, -y, -z'
 'x-1/2, -y-1/2, z-1/2'

_cell_length_a    11.8293
_cell_length_b    10.5512
_cell_length_c    21.6318
_cell_angle_alpha  90.000
_cell_angle_beta  100.203
_cell_angle_gamma  90.000

_shelx_F_squared_multiplier     1.000

loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0      772.37      856.47      28.20 o
   4   0   0     1445.15     1466.80      39.55 o
   6   0   0     1130.79     1097.08      30.42 o
   8   0   0     1347.13     1490.27      55.41 o
  10   0   0     3275.01     3545.64     154.91 o
  12   0   0       48.20       60.50       4.56 o
  14   0   0       79.87       63.02       7.91 o
   2   1   0     2093.70     1978.83      47.36 o
   3   1   0    33795.10    34884.29    1287.71 o
   4   1   0     2298.16     2035.72      38.24 o
   5   1   0        9.73       36.06       5.59 o
   6   1   0      449.80      506.89      11.92 o
   7   1   0        1.81        7.91       5.59 o
   8   1   0       43.36       28.81       6.79 o
   9   1   0       64.18       48.51       6.02 o
  10   1   0     1412.22     1628.54      45.96 o
  11   1   0      242.68      279.96       9.70 o
  12   1   0       14.96       10.52       3.84 o
  13   1   0       16.87       15.74       4.56 o
  14   1   0       16.46        7.91       7.91 o
  15   1   0        0.00        3.95       5.59 o
   0   2   0     2443.71     2679.14      61.27 o
   1   2   0    23397.80    23770.80     846.30 o
   2   2   0    20572.37    19502.51     520.01 o
   3   2   0     8856.88     8282.53     169.57 o
```

Key terminology – derived data

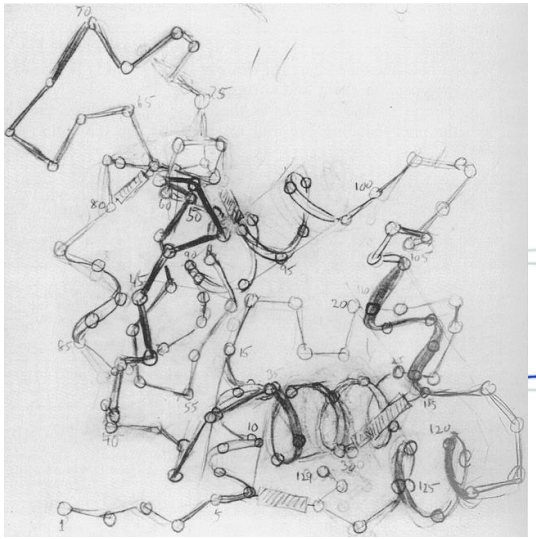
- 1929



Water Ice

Barnes, W. H. (1929). The Crystal Structure of Ice between 0º C and -183 º C. *Proc. R. Soc. London A* 125, 670-693.

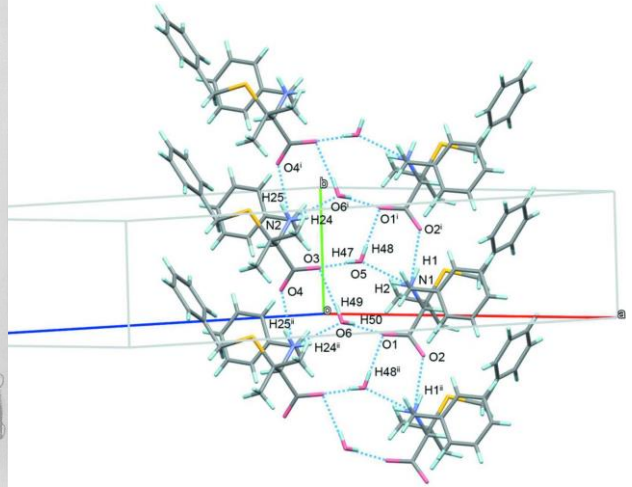
- 1965



Lysozyme

Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C., Phillips, D. C. & Sarma, V. R. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Ångstrom resolution. *Nature* **206**, 757-761. [Figure from Chapter 25.1 of *International Tables for Crystallography Volume F. Crystallography of biological macromolecules*]

- 2014

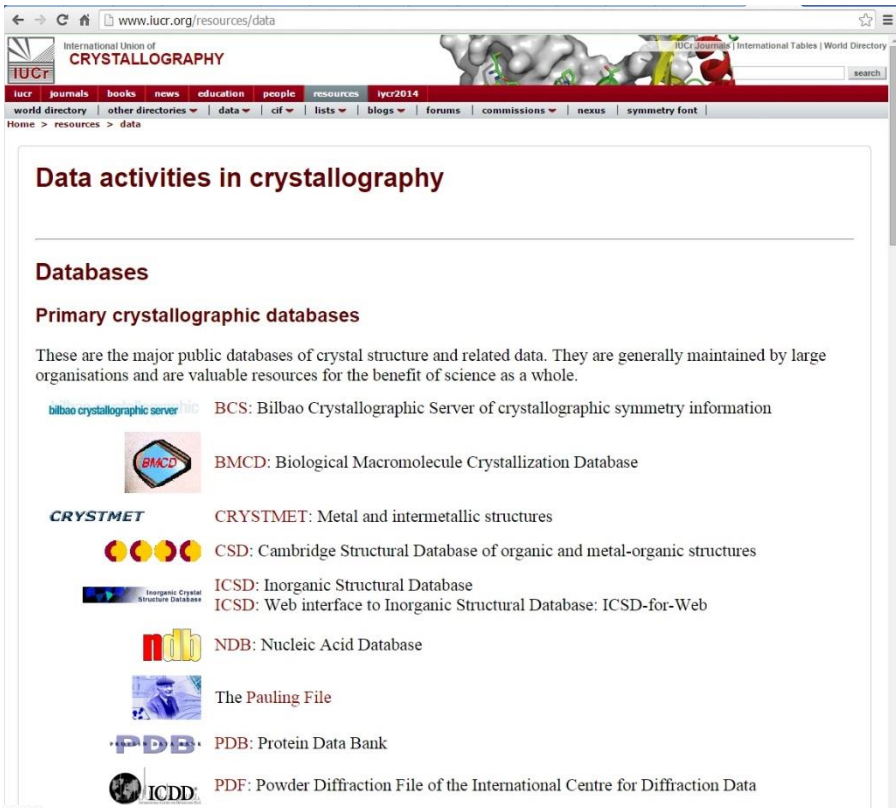


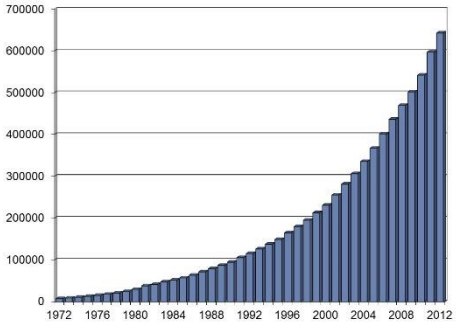
A penicillamine hydrate

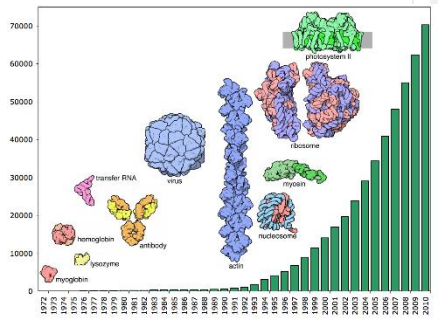
Yoshinari, N. & Konno, T. (2014). Crystal structure of *S,N*-dibenzyl-D-penicillamine monohydrate. *Acta Cryst.* E**70**, o1209

Benefits of retaining derived data

- Scientific record
- Database-driven discovery
- Protein-ligand interactions
- New pathways to synthesis, manufacturing, energetics…
- Identification/indexing (*e.g.* forensic science)



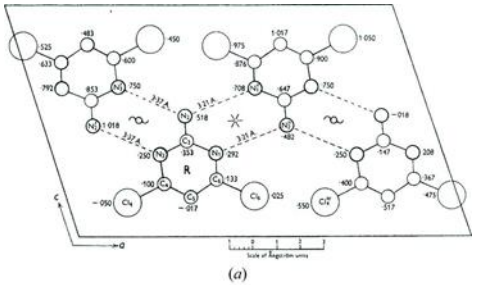




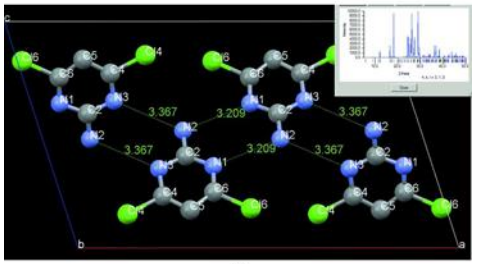
Derived data archived in CIF format

scripts.iucr.org/cgi-bin/sendcif?is5378sup1

```
_refine_ls_wR_factor_ref     0.0934
_refine_ls_goodness_of_fit_ref     1.063
_refine_ls_restrained_S_all     1.063
_refine_ls_number_reflns     7694
_refine_ls_number_parameters     461
_refine_ls_number_restraints     1
_refine_ls_hydrogen_treatment     mixed
_refine_ls_weighting_scheme     calc
_refine_ls_weighting_details
          'w=1/[\s^2^(Fo^2^)+(0.0482P)^2^+1.2220P] where P=(Fo^2^+2Fc^2^)/3'
_refine_ls_shift/su_max     0.000
_refine_ls_shift/su_mean     0.000
_refine_diff_density_max     0.435
_refine_diff_density_min     -0.252
_refine_ls_extinction_method     none
_refine_ls_extinction_coef     .
_refine_ls_abs_structure_details
;
Flack <i>x</i> determined using 2566 quotients
[(<i>I</i>^+^)-(<i>I</i>^-^)]/[(<i>I</i>^+^)+(<i>I</i>^-^)]
(Parsons <i>et al.</i>,  2013)
;
_refine_ls_abs_structure_Flack     0.03(2)
loop_
    _atom_type_symbol
    _atom_type_description
    _atom_type_scat_dispersion_real
    _atom_type_scat_dispersion_imag
    _atom_type_scat_source
    'C' 'C' 0.0033 0.0016
                          'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
    'H' 'H' 0.0000 0.0000
                          'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
    'O' 'O' 0.0106 0.0060
                          'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
    'N' 'N' 0.0061 0.0033
                          'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
    'S' 'S' 0.1246 0.1234
                          'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'

loop_
    _atom_site_type_symbol
    _atom_site_label
    _atom_site_fract_x
    _atom_site_fract_y
    _atom_site_fract_z
    _atom_site_U_iso_or_equiv
    _atom_site_adp_type
    _atom_site_calc_flag
    _atom_site_occupancy
    _atom_site_disorder_assembly
    _atom_site_disorder_group
    S S1 0.54021(3) 0.37960(9) 0.07484(2) 0.02754(16) Uani d 1 . .
    O O1 0.53088(10) -0.0974(3) 0.19789(6) 0.0392(5) Uani d 1 . .
    O O2 0.60252(10) -0.2717(3) 0.16081(6) 0.0349(5) Uani d 1 . .
    N N1 0.58790(11) 0.2829(3) 0.17415(7) 0.0240(4) Uani d 1 . .
    H H1 0.5905(13) 0.407(5) 0.1604(9) 0.029 Uiso d 1 . .
    H H2 0.5508(14) 0.309(5) 0.1889(9) 0.029 Uiso d 1 . .
    C C1 0.52331(12) 0.1335(4) 0.10419(8) 0.0244(5) Uani d 1 . .
    C C2 0.58099(12) 0.0945(4) 0.14360(8) 0.0234(5) Uani d 1 . .
    H H3 0.6245 0.0792 0.1314 0.028 Uiso calc 1 . .
    C C3 0.57073(12) -0.1113(4) 0.16993(8) 0.0260(5) Uani d 1 . .
    C C4 0.51755(14) -0.0596(4) 0.07331(9) 0.0347(6) Uani d 1 . .
    H H4 0.5008 -0.1831 0.0882 0.052 Uiso calc 1 . .
    H H5 0.5623 -0.0930 0.0655 0.052 Uiso calc 1 . .
    H H6 0.4859 -0.0268 0.0464 0.052 Uiso calc 1 . .
    C C5 0.45430(13) 0.1776(5) 0.11894(10) 0.0356(6) Uani d 1 . .
    H H7 0.4435 0.0616 0.1382 0.053 Uiso calc 1 . .
    H H8 0.4191 0.1858 0.0930 0.053 Uiso calc 1 . .
    H H9 0.4563 0.3135 0.1350 0.053 Uiso calc 1 . .
    C C6 0.6145(2) 0.3154(6) 0.04975(14) 0.0591(10) Uani d 1 . .
    H H10 0.6489 0.2435 0.0716 0.071 Uiso calc 1 . .
    H H11 0.6019 0.2172 0.0245 0.071 Uiso calc 1 . .
```



(a)



(b)

```
Alert level A
ATOM007_ALERT_1_A  _atom_site_aniso_label is missing
                   Unique label identifying the atom site.
GEOM001_ALERT_1_A  _geom_bond_atom_site_label_1 is missing
                   Label identifying the atom site 1.
GEOM003_ALERT_1_A  _geom_bond_distance is missing
                   Distance between atom sites 1 and 2.
GEOM004_ALERT_1_A  _geom_angle_atom_site_label_2 is missing
                   Label identifying the atom site 2.
PLAT027_ALERT_3_A  _diffrn_reflns_theta_full (too) Low .............  0.00 Deg.
PLAT029_ALERT_3_A  _diffrn_measured_fraction_theta_full Low .......  0.00
PLAT043_ALERT_1_A  Check Reported Molecular Weight .................  0.00
PLAT047_ALERT_1_A  Sum Formula Not Given ...........................  ?
PLAT130_ALERT_1_A  Volume as Calculated Differs from that Given ...  0.00 Ang**3
PLAT197_ALERT_1_A  Missing _cell_measurement_temperature ..........  ?

Alert level C
PLAT040_ALERT_1_C  No H-atoms in this Carbon Containing Compound ..  ?
PLAT128_ALERT_4_C  Non-standard setting of Space group P21/c  ....  P21/a
PLAT141_ALERT_1_C  No su (esd) Given on Volume .....................  ?
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  H1
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  C2
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  H2
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  H3
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  C4
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  Cl4
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  C5
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  C6
PLAT141_ALERT_4_C  Missing or Zero su (esd) on x-coordinate for ...  Cl6
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  H1
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  C2
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  H2
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  H3
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  C4
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  Cl4
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  C5
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  C6
PLAT142_ALERT_4_C  Missing or Zero su (esd) on y-coordinate for ...  Cl6
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  H1
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  C2
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  H2
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  H3
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  C4
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  Cl4
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  C5
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  C6
PLAT143_ALERT_4_C  Missing or Zero su (esd) on z-coordinate for ...  Cl6

 10 ALERT level A = In general: serious problem
  0 ALERT level B = Potentially serious problem
 30 ALERT level C = Check and explain
  0 ALERT level G = General alerts; check

 10 ALERT type 1 CIF construction/syntax error, inconsistent or missing data
  0 ALERT type 2 Indicator that the structure model may be wrong or deficient
  2 ALERT type 3 Indicator that the structure quality may be low
 28 ALERT type 4 Improvement, methodology, query or suggestion
  0 ALERT type 5 Informative message, check
```

(c)

Benefits of retaining processed data

- Structure validation
- Re-refinement
- Systematic bias, methods development
- Guard against structures associated with incorrect data sets

journals.iucr.org/b/issues/2002/01/00/bm0047/index.html

Structural Science
Crystal Engineering
and Materials

contents of issue
search
subscribe
help

navigate
pdf
supplementary files
next in issue

research papers

SHARE highlight terms

Structural Science

Volume 58
Part 1
Pages 62-77
February 2002

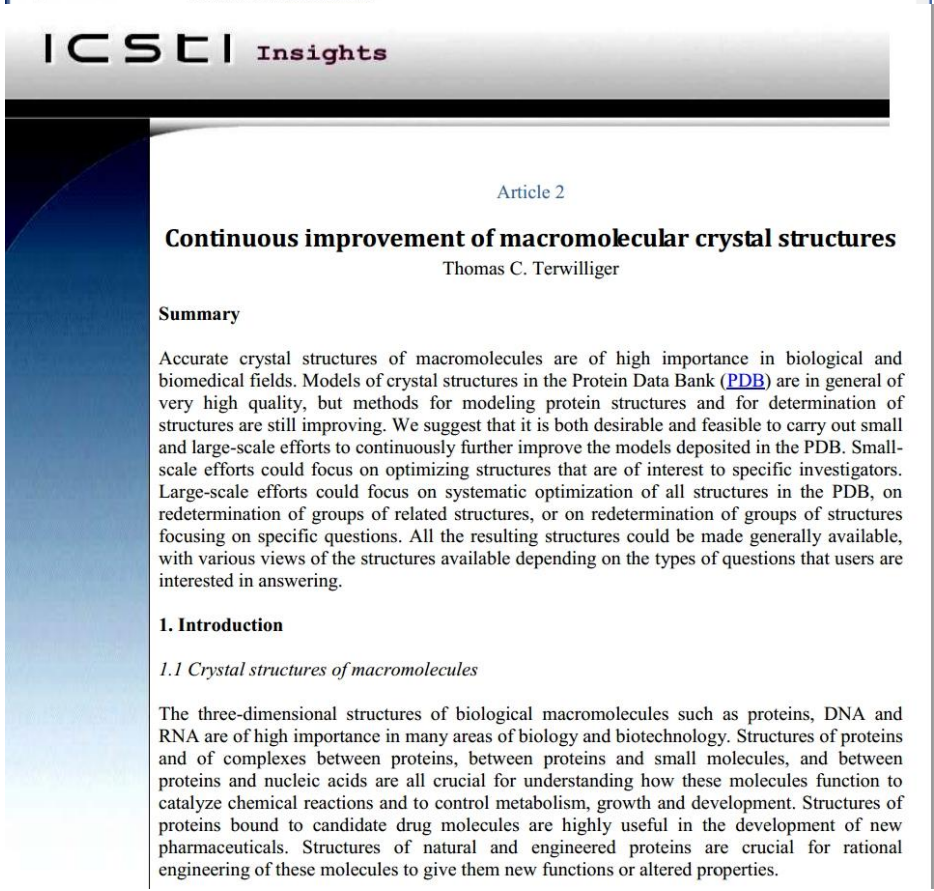
Received 2 June 2001
Accepted 12 October 2001

Some 60 new space-group corrections

Richard E. Marsh,[a] Moshe Kapon,[b] Shengzhi Hu[c] and Frank H. Herbstein[b]*

[a]The Beckman Institute, California Institute of Technology, Pasadena, CA 91125, USA, [b]Department of Chemistry, Technion-Israel Institute of Technology, Haifa 32000, Israel, and [c]Department of Chemistry, Xiamen University, Xiamen, Peoples' Republic of China
Correspondence e-mail: chr03fk@tx.technion.ac.il

CrossMark



ICSTI Insights

Article 2

Continuous improvement of macromolecular crystal structures

Thomas C. Terwilliger

Summary

Accurate crystal structures of macromolecules are of high importance in biological and biomedical fields. Models of crystal structures in the Protein Data Bank (PDB) are in general of very high quality, but methods for modeling protein structures and for determination of structures are still improving. We suggest that it is both desirable and feasible to carry out small and large-scale efforts to continuously further improve the models deposited in the PDB. Small-scale efforts could focus on optimizing structures that are of interest to specific investigators. Large-scale efforts could focus on systematic optimization of all structures in the PDB, on redetermination of groups of related structures, or on redetermination of groups of structures focusing on specific questions. All the resulting structures could be made generally available, with various views of the structures available depending on the types of questions that users are interested in answering.

1. Introduction

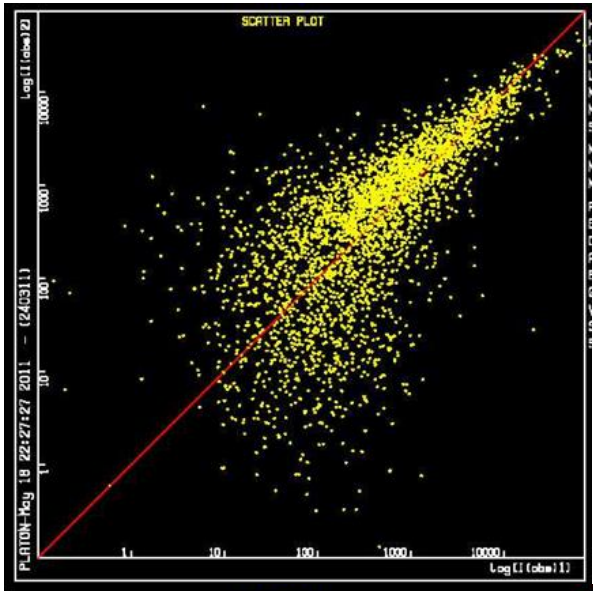
1.1 Crystal structures of macromolecules

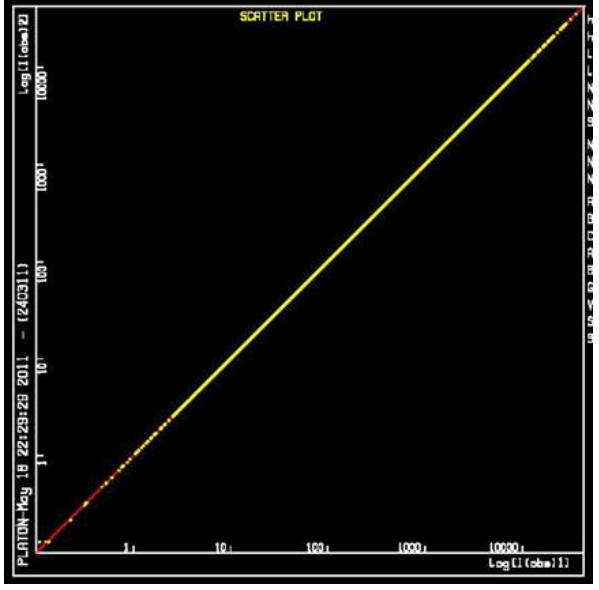
The three-dimensional structures of biological macromolecules such as proteins, DNA and RNA are of high importance in many areas of biology and biotechnology. Structures of proteins and of complexes between proteins, between proteins and small molecules, and between proteins and nucleic acids are all crucial for understanding how these molecules function to catalyze chemical reactions and to control metabolism, growth and development. Structures of proteins bound to candidate drug molecules are highly useful in the development of new pharmaceuticals. Structures of natural and engineered proteins are crucial for rational engineering of these molecules to give them new functions or altered properties.

Processed data archived in CIF format

journals.iucr.org/e/issues/2014/11/00/is5378/is5378Isup2.hkl

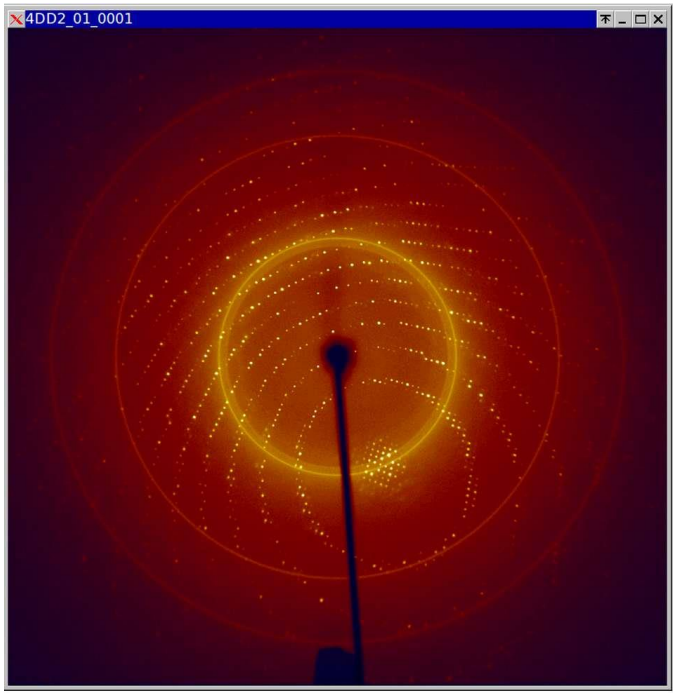
```
#
# h,k,l, Fc-squared, Fo-squared, sigma(Fo-squared), status flag
#
data_I
_shelx_title ' 1'
_shelx_refln_list_code          4
_shelx_F_calc_maximum      388.47
_exptl_crystal_F_000      1488.00
_reflns_d_resolution_high  0.7700

loop_
 _space_group_symop_operation_xyz
 'x, y, z'
 '-x, y, -z'
 'x+1/2, y+1/2, z'
 '-x+1/2, y+1/2, -z'

_cell_length_a    19.9301
_cell_length_b     6.2500
_cell_length_c    30.6453
_cell_angle_alpha  90.000
_cell_angle_beta   98.715
_cell_angle_gamma  90.000

_shelx_F_squared_multiplier     1.000

loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   0  -8   0         239.32         195.13          83.01 o
   2  -8   0          26.69          83.64          61.48 o
   4  -8   0          56.89          97.69          34.05 o
   1  -7   0         112.86          88.84          29.00 o
   3  -7   0         682.19         602.61          37.64 o
   5  -7   0         633.07         701.40          39.03 o
   7  -7   0         217.84         236.71          31.28 o
   9  -7   0         214.09         314.21          45.05 o
  11  -7   0         390.37         396.62          33.39 o
   0  -6   0         731.14         758.52          48.57 o
   2  -6   0         607.38         552.97          32.54 o
   4  -6   0         121.37         157.04          24.13 o
   6  -6   0         198.12         212.12          25.42 o
   8  -6   0         296.06         286.91          49.78 o
  10  -6   0         273.93         299.07          32.52 o
  12  -6   0         457.20         477.48          69.09 o
  14  -6   0         211.28         210.26          31.27 o
  16  -6   0         640.99         608.94          34.45 o
   1  -5   0        2288.34        2255.88          52.35 o
   3  -5   0         685.42         651.06          31.85 o
   5  -5   0          24.58          36.08          17.51 o
   7  -5   0        1008.38         911.86          30.59 o
   9  -5   0         906.06         852.23          43.59 o
  11  -5   0        2289.15        2200.00          52.99 o
  13  -5   0        1947.78        1866.15          48.14 o
  15  -5   0        1256.33        1208.63          42.74 o
  17  -5   0         544.92         577.60          27.83 o
  19  -5   0         356.67         355.18          50.48 o
   0  -4   0        2794.01        2551.43          71.20 o
   2  -4   0        4618.55        4739.23         103.09 o
   4  -4   0        8314.14        7536.70         155.97 o
   6  -4   0        2354.85        2200.28          49.94 o
   8  -4   0        1857.12        1850.00         105.21 o
  10  -4   0         524.79         495.91          23.30 o
  12  -4   0        2364.29        2267.69          46.76 o
  14  -4   0        1979.45        1865.69          50.91 o
  16  -4   0          58.45          53.41          15.44 o
  18  -4   0         165.45         157.99          17.89 o
  20  -4   0         317.37         333.75          27.15 o
```





Benefits of retaining raw data

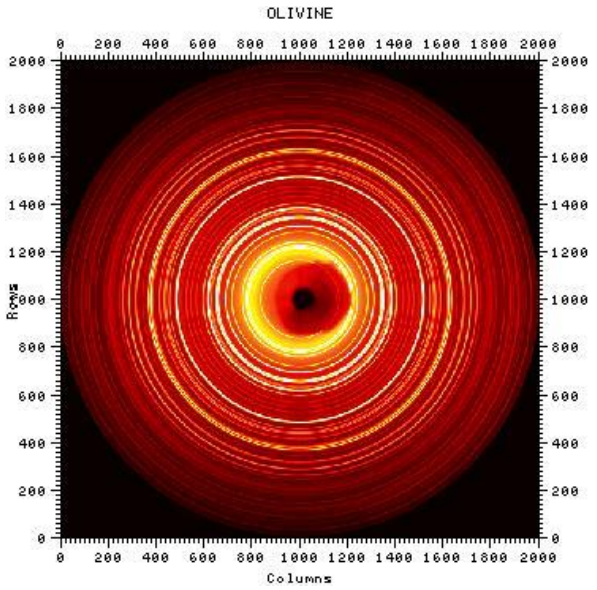
- Structure validation
- Re-refinement
- Systematic bias, methods development
- Guard against structures associated with incorrect data sets

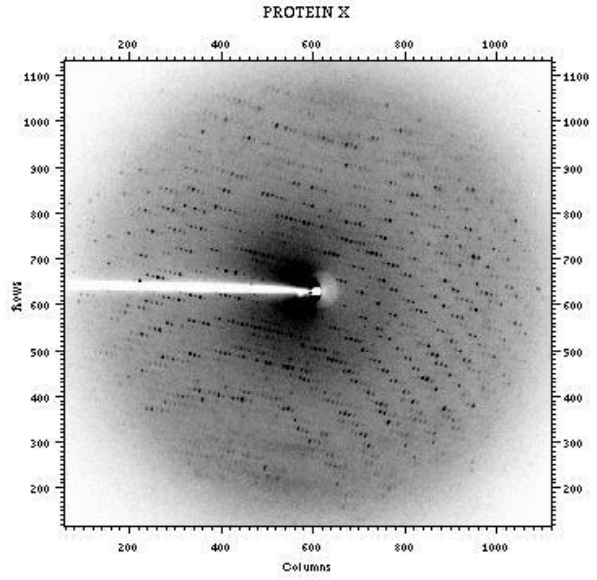


Raw data archived in CIF format

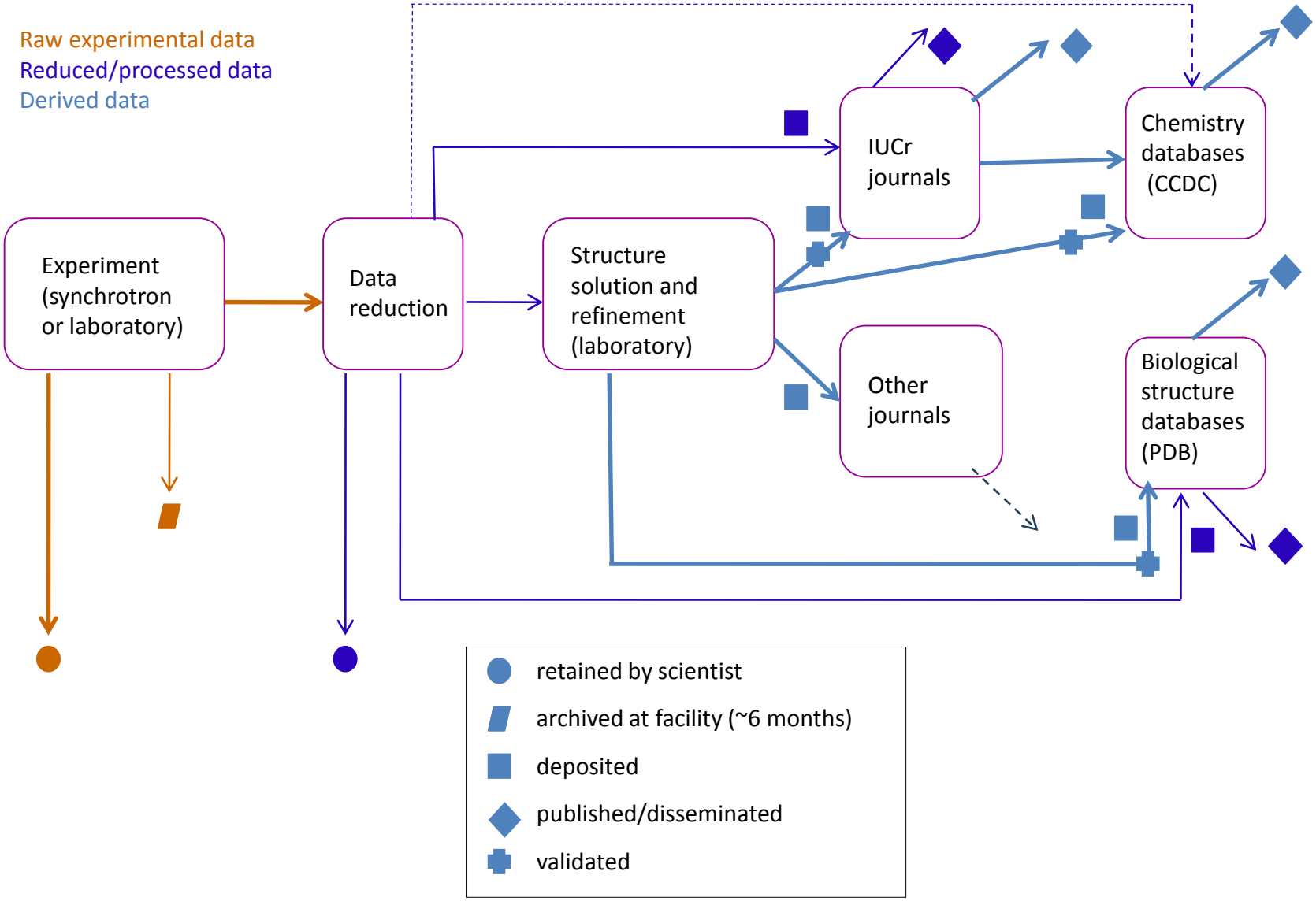
Example 2.3.4.1. *A multiple-block CBF with several images.*

```
###CBF: VERSION 1.0
# CBF file written by cbflib v0.6

# A comment cannot appear before the file identifier,
# but can appear anywhere else, except within the
# binary sections.

# Here the first data block starts
data_xxx

### ... various CIF tags and values here
###     but none that define array data items

# The "data_" identifier finishes the first data
# block and starts the second
data_yyy

### ... various CIF tags and values here including
###     ones that define array data items

loop_
_array_data.array_id
_array_data.binary_id
_array_data.data

image_1 1
;
--CIF-BINARY-FORMAT-SECTION--
Content-Type: application/octet-stream;
     conversions="x-CBF_PACKED"
Content-Transfer-Encoding: BINARY
X-Binary-Size: 3745758
X-Binary-ID: 1
X-Binary-Element-Type: "signed 32-bit integer"
Content-MD5: 1zsJjWPfol2GYl2V+QSXrw==

START_OF_BIN
<D5>^P<B8>P^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@ ...
[This is where the raw binary data would be – we can't print them here]

--CIF-BINARY-FORMAT-SECTION----
;
```





Data flow in crystallography



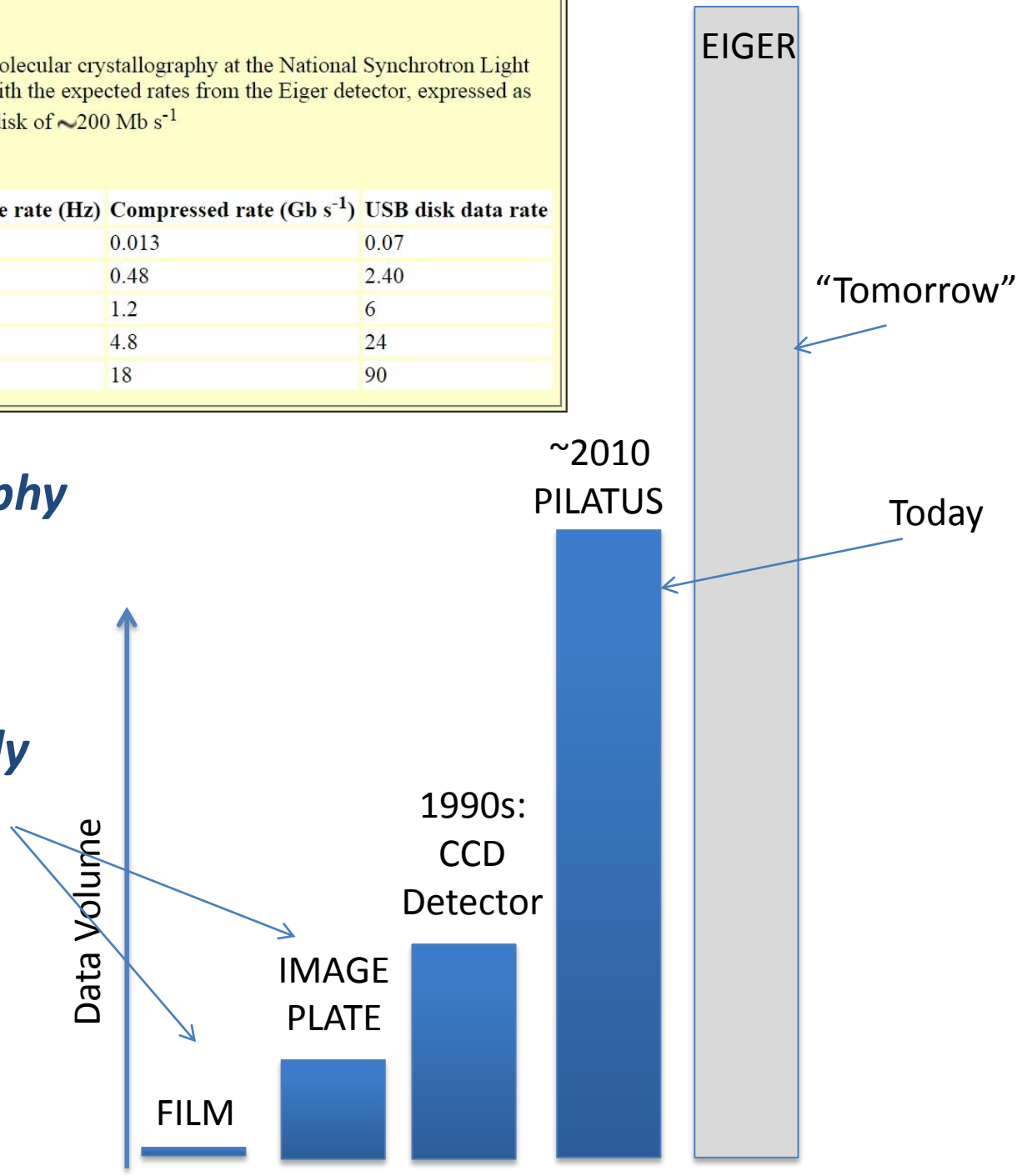
Modern data flows in biological crystallography are ***extreme***>>>>

Table 1
Typical sustained data rates for detectors used for macromolecular crystallography at the National Synchrotron Light Source and Diamond Light Source beamlines compared with the expected rates from the Eiger detector, expressed as multiples of the typical data rate for an inexpensive USB disk of ~200 Mb s^{-1}

From Bernstein *et al.* (2013).

Detector	Raw image size (MB)	Frame rate (Hz)	Compressed rate (Gb s^{-1})	USB disk data rate
ADSC Q315 (2 × 2 binned)	18	0.37	0.013	0.07
PILATUS 6M	24	10	0.48	2.40
PILATUS 6M-F (fast)	24	25	1.2	6
PILATUS3 6M	24	100	4.8	24
Eiger 16M	72	125	18	90

Biological Crystallography Labs at Synchrotron Radiation Sources; data flows have increased dramatically since the 1980s



MANCHESTER 1824
The University of Manchester
Manchester Institute of Biotechnology

Modern data flows in biological crystallography are *extreme*

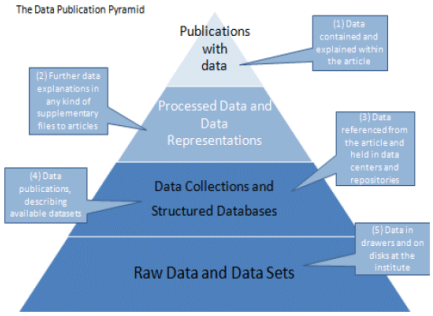
Would keeping all these raw data be “worth the pain”?

A group of 4 recent articles in *Acta Crystallographica Section D: Biological Crystallography* explains why keeping raw data is a ***natural next step*** for crystallography ….

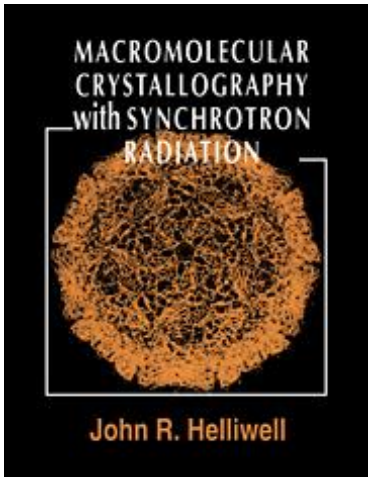
Acta Crystallographica Section D: Biological Crystallography
Volume **70**, Part 10 (October 2014): special section on
Diffraction Data Deposition

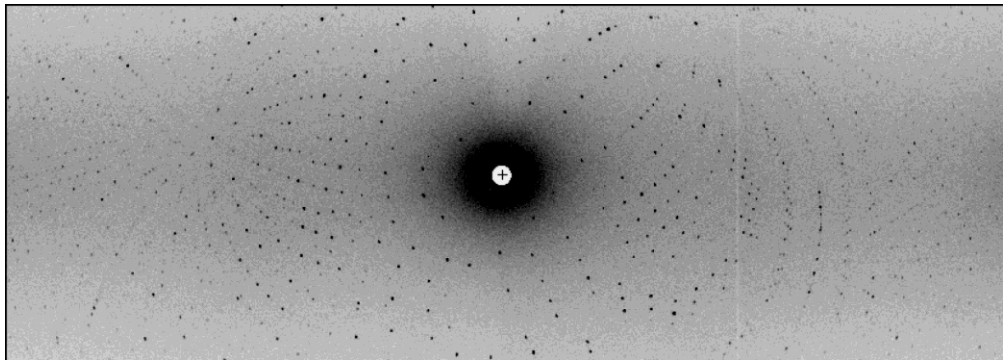
- Terwilliger, T. C. (2014). **Archiving raw crystallographic data.** *Acta Cryst.* D**70**, 2500-2501 *(Editorial)*
- Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). **Experiences with making diffraction image data available: what metadata do we need to archive?** *Acta Cryst.* D**70**, 2502-2509
- Meyer, G. R., Aragao, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). **Operation of the Australian Store.Synchrotron for macromolecular crystallography.** *Acta Cryst.* D**70**, 2510-2519.
- Guss, J. M. & McMahon, B. (2014). **How to make deposition of images a reality.** *Acta Cryst.* D**70**, 2520-2532
- Terwilliger, T. C. & Bricogne, G. (2014). **Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data.** *Acta Cryst.* D**70**, 2533-2543

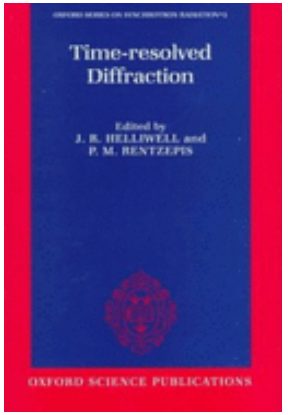
Raw diffraction images offer the opportunity of:-

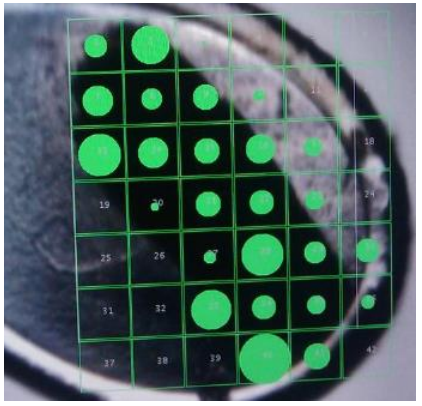


- *analysing data at higher resolution than used in the original work*
- *serving as benchmarks in developing improved methods of analysis*
- *checking the interpretation of the symmetries of the crystals*
- *analysing diffraction from multiple lattices present in the crystals*
- *analysing the diffuse scattering that reflects correlated motions or disorder of atoms in the crystals*


MACROMOLECULAR
CRYSTALLOGRAPHY
with SYNCHROTRON
RADIATION
John R. Helliwell

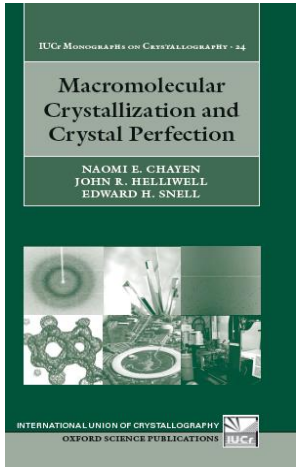



Time-resolved
Diffraction
Edited by
J. R. HELLIWELL and
P. M. RENTZEPIS
OXFORD SCIENCE PUBLICATIONS




diamond


Thank you


IUCr Monographs on Crystallography - 24
Macromolecular
Crystallization and
Crystal Perfection
NAOMI E. CHAYEN
JOHN R. HELLIWELL
EDWARD H. SNELL
INTERNATIONAL UNION OF CRYSTALLOGRAPHY
OXFORD SCIENCE PUBLICATIONS
IUCr