Inorganica Chimica Acta 470 (2018) 232–237

Contents lists available at ScienceDirect

Inorganica Chimica Acta

journal homepage: www.elsevier.com/locate/ica

Research paper

What makes a crystal structure report valid?

Anthony L. Spek

Crystal and Structural Chemistry, Bijvoet Centre for Biomolecular Research, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 23 March 2017
Accepted 18 April 2017
Available online 24 April 2017

Special Volume: Protagonists in Chemistry
Dedicated to Professor Carlo Mealli

Keywords:
Validation
CheckCIF
PLATON
SQUEEZE
MOF
Crystalline Sponge Method

ABSTRACT

Single crystal X-ray crystallography has developed into a unique, highly automated and accessible tool to obtain detailed information on molecular structures. Proper archival makes that referees, readers and users of the results of reported crystal structures no longer need to depend solely on the expertise of the analyst, often a non-professional crystallographer, who did the reported study. Deposited computer readable data should allow for an independent structure analysis, validation of the author's interpretation of the experimental data and use of those data for follow-up research. This paper summarises what is needed for proper validation and archival. The difference between *valid* and *value* is discussed. As an example, the deposited data associated with the molecular structure determination of a guest molecule soaked into a MOF, based on the *Crystalline Sponge Method*, are analysed.

1. Introduction

The majority of papers published in chemical journals such as Inorganica Chemica Acta, Inorganic Chemistry and Organometallics include one or more crystal structure reports. In many cases those structures mainly serve as 'solid proof' of the identity of a compound in the context of the reported chemical research: *Seeing is Believing*. Many of those crystal structure determinations are nowadays, once suitable crystals are obtained, routine in the hands of experienced analysts. The reported structures do not necessarily offer significant new chemical or crystallographic insight on their own. For that reason, only limited details beyond a 3D representation and a footnote with selected data normally appear in print along with a deposition reference for more details. Often, the main added value of a structure determination lies in their subsequent inclusion in the Cambridge Structural Database (CSD) [1] that offers a rich source of data for all kinds of comparative, statistical and follow-up research. For the latter, quality, reliability and completeness of the deposited data is essential. It is important that all reported crystal structures are well documented and validated. Sufficient information should be made available to allow for an independent structure analysis with the archived data. Sometimes, those data are unique such as meta-stable polymorphs or difficult and costly to obtain again from scratch.

For many purposes, the availability of the set of coordinates of the atoms in a molecule is sufficient for detailed geometry calculations and the preparation of a 3D illustration. The CSD, maintained by the Cambridge Crystallographic Data Centre (CCDC [1]), makes those data for published structures readily available along with molecular graphics and analysis tools. More details on a structure determination can generally be found in the archived and freely available CIF file, which is readable both by humans and by computers.

The CIF [2] standard for data exchange and archival was pioneered by the International Union of Crystallography (IUCr) [3]. This standard allows for automatic structure validation, through the IUCr/checkCIF [4,5] webserver, of the archived data in a CIF for completeness, consistency and quality against common standards. In its original implementation, where mainly the refinement results were archived, a CIF effectively only documented the author's interpretation of the experimental diffraction data. With that information, interpretation errors are often difficult to detect, prove and investigate. The current standard is therefor to also archive the refinement details and the unmerged reflection data into a deposited CIF. That allows referees and readers to do their own analysis of the experimental data when interpretation questions arise, in particular when unusual results are claimed or

E-mail address: a.l.spek@uu.nl

http://dx.doi.org/10.1016/j.ica.2017.04.036
0020-1693/

spotted by experts. Current versions of structure refinement packages such as SHELXL [6], Olex2 [7] will create by default those extended CIF files. In the future, deposition of the original diffraction images may become an option/standard as well [8]. With those images it should be possible to search for diffraction effects that were not included in the data reduction step of the analysis.

Inadequate interpretation and handling of the diffraction data by analysts with no formal training can be a problem. Validation software offers a tool to alert for issues that need to be addressed before publication. Common problems and pitfalls are missassigned atom types, too many or too few hydrogen atoms, disorder, missed twinning and missed higher symmetry, all possibly leading into false reported chemistry. Sometimes erroneous interpretations lead to false concepts such as the illusory '*bond-stretch isomerism*' [9], i.e. bonds with a double energy minimum, that later was shown to be caused by substitutional disorder with a contaminant. Thanks to experts such as Carlo Mealli, false reported structures are eventually spotted, investigated and corrected [10]. In this journal, Clemente [11] has reported on necessary space group changes and their chemical consequences.

This paper discusses various structure validation issues, illustrated with an analysis of a *Crystalline Sponge Method* based structure report as an example.

2. Validation tools

Various readily available structure validation tools are used by authors, referees, journal editors and readers to evaluate structure reports. Those tools are not completely independent but allow looking at a structure report from different perspectives.

2.1. The R-value

A popular practice is to look for low R-values. The assumption here is that the quality and correctness of a structure can be measured with a single number. The premise is that a low value of the disagreement factor, R, between the observed and calculated structure factors, say R < 5%, can be taken as an indication for a good structure. The problem with that is that, e.g. in the case of a Uranium based metal-organic compound, the scattering contribution due to the heavy Uranium atom can be so large that a wrong interpretation in the weaker scattering organic part of the structure will have only a minor effect on the R-value. Wrong atom type assignments and missing or to many hydrogen atoms in a structure model may go unnoticed for that reason.

2.2. ORTEP illustration

An ORTEP plot [12] provides a 3D graphical summary of most of the refined model parameters. In particular the shape, direction and size of the ellipsoids can visually point to unresolved problems. The reason for extreme disc or cigar shaped ellipsoids should be investigated and acted upon. Common reasons are (substitutional) disorder, poor data and symmetry related issues. A missed centre of inversion may show up as unequal but chemically identical bond distances and perpendicular main axes of the displacement parameters of inversion related atoms (see example 2 in Ref. [5]). The problem is that signals for an issue with a structure can be hidden with suitable constraints and restraints on the coordinates and displacement parameters, U^{ij}'s, at the cost of a higher R-value that can be blamed to 'poor data'. Fig. 1 provides an example [13], deposition code CCDC 1470206, of a nice ORTEP. The naphthyl moiety in this figure serves as a reference for the ORTEP plots shown in Fig. 5.

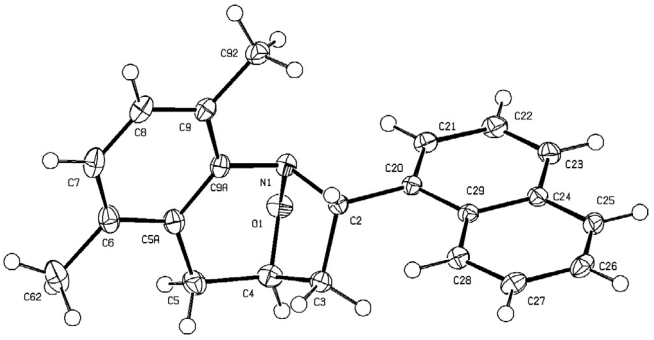

Fig. 1. A 100 K quality structure ORTEP illustration. Only one of the two closely identical but crystallographically independent molecules is shown. Displacement ellipsoids are drawn at the 50% probability level. The ellipsoids of the naphthyl moiety have to be compared with those in Fig. 5.

2.3. Refinement details

This involves checking details such as to whether the least-squares refinement converged and whether constraints & restraints were used. Constraints and restraints may hide problems with a structure. Their use may indicate poor reflection data, a poor observed data over parameter ratio and/or disorder. Detailed interpretation and discussion of intra- and inter-molecular geometry may not be valid in such a case. Also unusual values of the refined values of the reflection weight model should be explained. It is also relevant to investigate the results of the *Analysis-of-Variance* statistics and outlier reflections being either measurement errors or inadequacies in the refinement model. Hydrogen atoms on heteroatoms such as N and O should be refined to prove their validity.

2.4. Difference electron density map

The final difference density map should be essentially clean apart from low-level noise excursions due to experimental and model errors. Such a map shows that the electron density map as calculated with the refined model parameters matches the one calculated with the observed reflection data. As an example, a

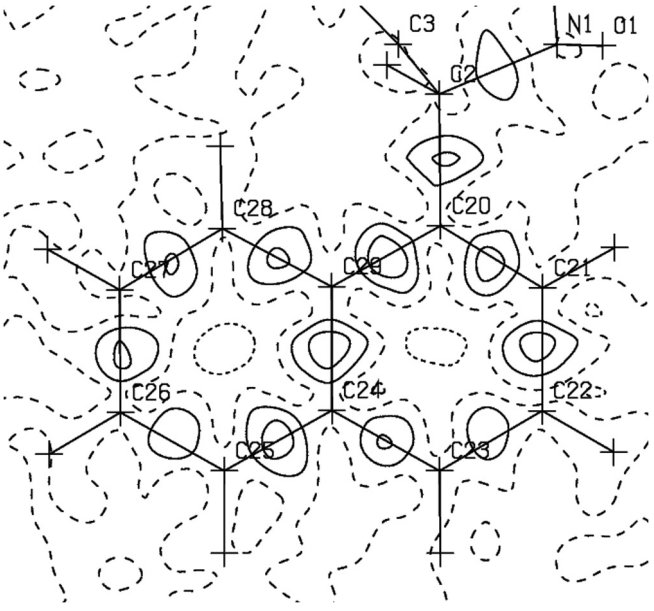

Fig. 2. Contoured section of the electron density difference map for the structure shown in Fig. 1. The residual density maxima on bonds are indicative of a good quality structure. Contour lines, solid for positive values and dashed for zero or negative values, are drawn with 0.1 e/Å^3 increments.

contoured section of the electron difference density map defined by the planar naphthyl moiety of the structure depicted in Fig. 1 is shown in Fig. 2. A good structure will show residual density maxima on the ring bond centres. The reason for that is that the deviation from spherical atomic density due to bonding effects is generally not taken into account in the refined set of model parameters. There should be no significant density maxima on the atom sites. A positive density peak on an atom site may indicate that the assigned atom type needs to be changed into one with a higher atomic number (e.g. N into O). Similarly, a negative density maximum on such a site may indicate an assigned atom type with a too high atom number or a non-unit site occupation due to disorder. A positive density peak value near light atoms might indicate a missed hydrogen atom and a negative value may point to an erroneously included hydrogen atom. Density maxima and minima are also often found around heavy atoms, generally at distances in the order of 1 Å or less. In most cases they can be interpreted as absorption artefacts due to insufficient or incorrect correction for absorption. Other sources of residual density maxima are unaccounted for (disordered) solvent molecules, substitutional disorder and twinning.

2.5. IUCr/checkCIF

IUCr/checkCIF is for a large part based on the structure validation tool available in the PLATON program [4]. Tests are done for completeness and consistency of the data, proper procedures and issues to be addressed such as symmetry problems, unaccounted for solvent accessible voids in a structure, unusual refined parameter values. ALERTS are generated with levels A, B or C. In addition, G-ALERTS will report on issues that are not necessarily errors but worth to investigate and/or discuss. Examples are messages about the special constraints and restraints applied to the model parameters. Potentially missed symmetry ALERTS are purely based on symmetry relations between atomic coordinates. Reflection data are needed for a detailed analysis of such ALERTS.

A low R(int) value, the averaging index of multiple and symmetry related reflection intensity measurement, can be an indicator for a good data set.

A normal Probability Plot tests whether differences between observed and calculated structure factors are normally (i.e. Gaussian) distributed. A generally linear plot is expected. A large deviation, in particular in the tails, generally points to data and/or model problems.

2.6. CSD search

The CSD can be used to search for precedents for a supposedly unusual feature in a structure. The knowledge-based library MOGUL [14] that comes with the CSD may also be helpful for a comparison of the geometry of fragments in the structure at hand with the geometry of similar fragments in archived structures.

2.7. Experience and chemical insight

Not all interpretation errors of the experimental data can be detected automatically. Experience with known pitfalls and in particular chemical insight are still very important. IUCr/checkCIF sends out G-level Alerts that call for such expertise.

3. Common issues

Common problems are disorder, missed twinning signs and pseudo-symmetry. It is not always clear which one applies. 'Disorder' might well be an artefact of a twinning or a symmetry problem. Severely disordered solvents of crystallization are easily overlooked, in particular when one relies solely on the residual density peak list as reported by the refinement program used. The peak search software generally searches only for isolated density minima and maxima and might therefor overlook density ridges in incommensurately filled solvent channels. Weights applied to the reflections in the least squares refinement are often optimized to reach a Figure-of-Merit value (S value) near 1.0. Failure to reach a value close to 1.0 or unusual weight parameter values may point to unresolved issues.

4. An illustrative validation example

Not all compounds of interest crystallize readily. Often, a multitude of solvents and solvent mixtures have to be tried before crystals, suitable for an X-ray study, are obtained. Sometimes only chemical modification of the molecule of interest such as making the target compound into a salt will do, so might co-crystallization with a hydrogen bond acceptor such a triphenylphosphine oxide [15].

In 2013, Fujita et al. [16] introduced an interesting new approach to obtain structural information on difficult-to-crystallize compounds: the *Crystalline Sponge Method*. The basic idea of that technique is simple: use a crystal with suitable channels filled with a solvent that is easily replaced by soaking that crystal with the molecule of interest and solve and refine the resulting crystal structure. Metal-organic-framework structures, MOF's, naturally present themselves for this approach. Early proof-of-concept experiments [16] were done with the MOF framework $[(ZnI_2)_3(\text{tris(4-pyridyl)triazine})_2]_n$. That framework turned out to have several disadvantages such as disorder in the metal coordination sphere that needs to be addressed with a disorder model and the presence of an unnecessary strong coordinating Iodine scatterer. A subsequent search in the Cambridge Structural Database (CSD) suggested a more promising MOF candidate: [CuBr(benzene-1,3,5-triyl-triisonicotinate)]$_n$. Its 3D framework (Fig. 3) contains two approximately equally sized but crystallographically independent infinite channels, A & B, with a periodically repeated solvent accessible volume of ~700 Å^3 each. The monoclinic unit cell includes four of those channels covering ~40% of the unit cell volume. Fujita et al. [17] published as an example the sponge structure of 1-acetonaphthone soaked into this new 3D MOF along with the associated refinement and reflection data (CCDC deposition code 1511768). This allows us to illustrate its structure validation, to investigate the quality of the MOF framework and to

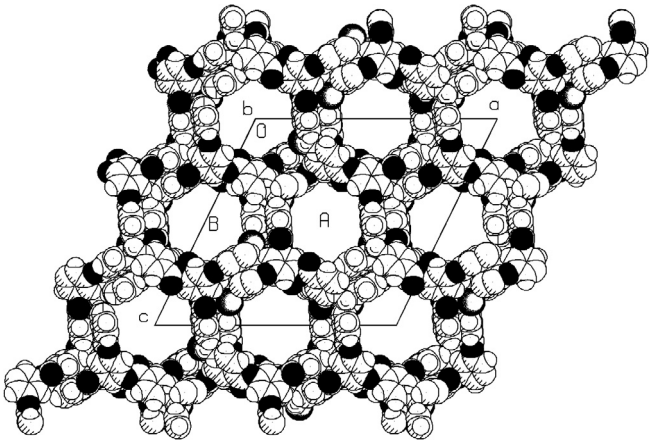

Fig. 3. The framework structure of the MOF structure [CuBr(benzene-1,3,5-triyl-triisonicotinate)]$_n$ with space group $P2_1/c$. Atoms are drawn with their van der Waals radii. The framework features two (A & B) crystallographically independent infinite channels. Each channel includes a mixture of the guest molecule and the solvent CS_2 molecules.

evaluate the quality of the sought for molecular geometry of the embedded 1-acetonaphthone guest molecule. Of particular interest is also to investigate the achieved level of saturation with the guest molecule in the two similar but crystallographically independent channels and, when not 100%, whether there are traces of solvent molecules at the site of the guest molecule when not taken by the guest molecule of interest.

4.1. checkCIF report

There are no potentially serious level A or B Alerts. However there are a number of noticeable level C and type G Alerts that need attention. The low R(int) value 0.018 suggests a good data set. The largest residual density maxima up to 1.74 e/Å^3 are near the CS_2 solvent molecules, suggesting unaccounted for solvent disorder in that part of the structure. The second SHELXL style optimized weighting parameter value of 19.01 is relatively high. An unusually large number of reflections, 335, is missing below sin $(\theta)/\lambda = 0.6$. Various constraints and restraints are used in the refinement on bonds and displacement parameters. There are no residual density maxima on C–C bonds. The Normal Probability Plot deviates significantly from linearity in the tails.

4.2. The MOF-framework

The quality of the new MOF framework can be investigated with the PLATON/SQUEEZE [18] tool. In this test, SQUEEZE takes care of the contribution of the diffracting content of the channels in a MOF in the least-squares refinement without the need to parameterize the channel content. For this purpose, all non-framework guest molecule and solvent atoms were removed from the refinement model. Doing so, unrestrained refinement of the framework parameters nicely converged at $R_1 = 0.0225$, $wR_2 = 0.0644$, $S = 1.085$. Residual density ranges nicely between −0.35 and 0.43 e/Å^3. The displacement ellipsoid plot of the MOF looks good and similar to that of the published structure. This result is taken as an indication that the reflection data are of high quality. Obviously, this approach does not provide a nicely refined model for the embedded guest molecules in the channel when those are the target of the study. What is shown with the SQUEEZE calculation is that both the A & B channels contain approximately the same integrated density electron count (i.e. 221 & 222 electrons). Also it is clear that both channels contain mixtures of the original CS_2 content and the target molecule as gleaned from the optimized difference density map obtained with the SQUEEZE tool.

4.3. Full model refinement

The authors of [17] refined a structure model with a partially occupied (s.o.f = 0.795(4)) target molecule site in the A channel and a fully occupied target molecule site (s.o.f. = 1.0) in the B channel, completed with partially occupied CS_2 solvent sites elsewhere in both channels. Refinement converged at $R_1 = 0.0529$, $wR_2 = 0.1818$, $S = 1.066$. The not explicitly reported residual density ranged between −1.26 and 1.74 e/Å^3, with the highest peaks near the CS_2 solvent molecules. From those values, in comparison to the residual density range achieved above (Section 4.2) with the SQUEEZE description of the channel content, it is clear that the reported refinement model is incomplete in accounting for all density in the A & B channels.

The guest molecule geometry in channel A was refined without restraints. The C—C bond distances in the naphthyl moiety deviate by −0.10 to 0.06 Å from the corresponding distances in the quality structure [13] depicted in Fig. 1. This large range contrasts unfavourably with the maximum bond distance difference of 0.005 Å between the two independent molecules in Ref. [13].

It should be noted that various restraints were used by the authors on the values of the displacement parameters of the guest molecule in the B channel in order to keep them within reasonable bounds. Of particular relevance is that the naphthyl moiety C–C bonds of that molecule were constrained to the same value of 1.390 Å where those values are expected to range from ∼1.37 to ∼1.44 Å.

The unequal population of the target molecules in channels A & B, in contrast to what is found with the SQUEEZE based refinement was puzzling and suggested a new refinement where the population value of the guest molecule in both sites was to be refined.

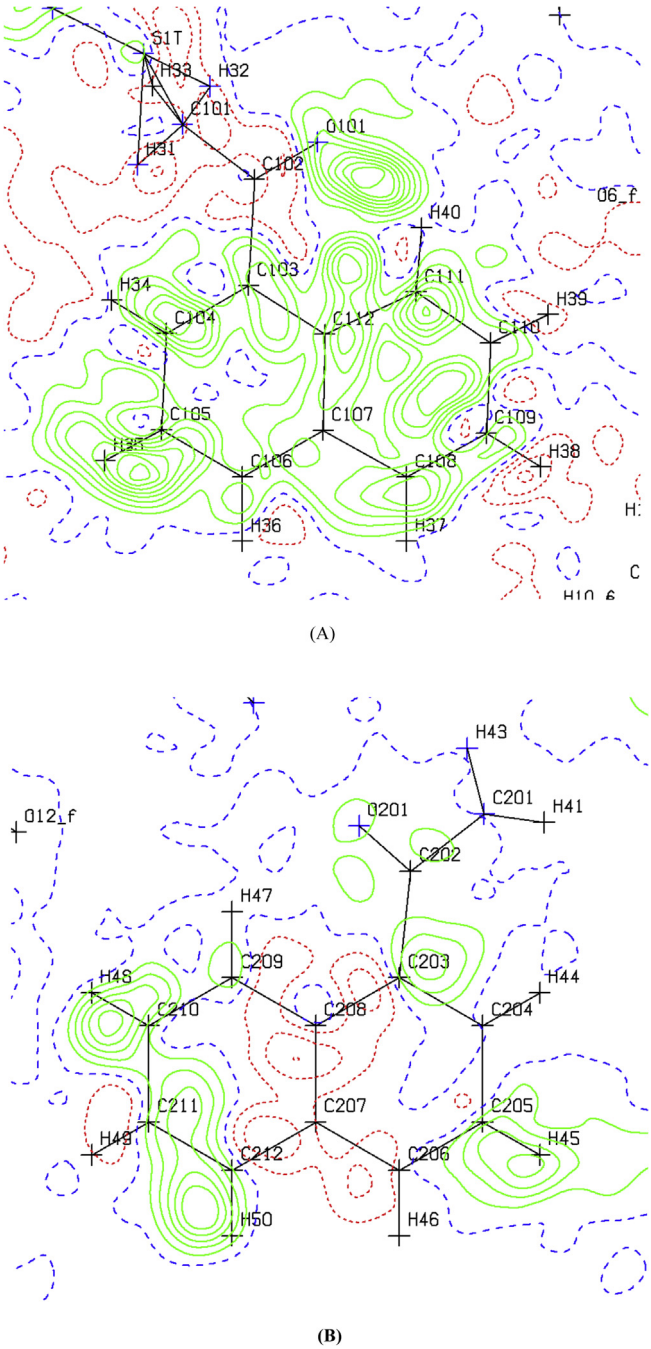

Fig. 4. Contoured difference density map sections for the two crystallographically independent guest molecules in channel A & B respectively. Both maps clearly illustrate the unaccounted for solvent molecules at the same lattice site. Contours, solid for positive values and dashed for zero or negative values, are drawn with 0.1 e/Å^3 increments.

That resulted in the now very similar population values of 0.793(4) for the guest molecule in the A channel and 0.769(5) for the guest molecule in the B channel, with improved R-values: $R_1 = 0.0509$, $wR_2 = 0.1705$, $S = 1.067$ and residual density excursions between −0.93 and 1.76 e/Å^3.

The contoured difference maps in Fig. 4, based on the above re-refinement, clearly show part of the reason for the high residual density excursions in the final residual density maps. It is obvious from those maps that part of the space not taken up in the sites that are partially occupied with the target molecule is taken up by CS_2 molecules, not taken into account in the refined parameter model. The ORTEP illustrations in Fig. 5 show the displacement parameters of the two guest molecules. The displayed ellipsoids are not of the quality as shown in Fig. 1 and larger than expected for a low temperature structure. Significant restraints were needed for the guest molecule in the B channel to compensate for the otherwise extremely poor geometry with C—C distances ranging from 1.12 to 1.67 Å when refined without constraints and restraints. The ellipsoids of the molecule in channel B, containing O201, show the effect of the applied SHELXL SIMU style restraint on the U_{ij}'s.

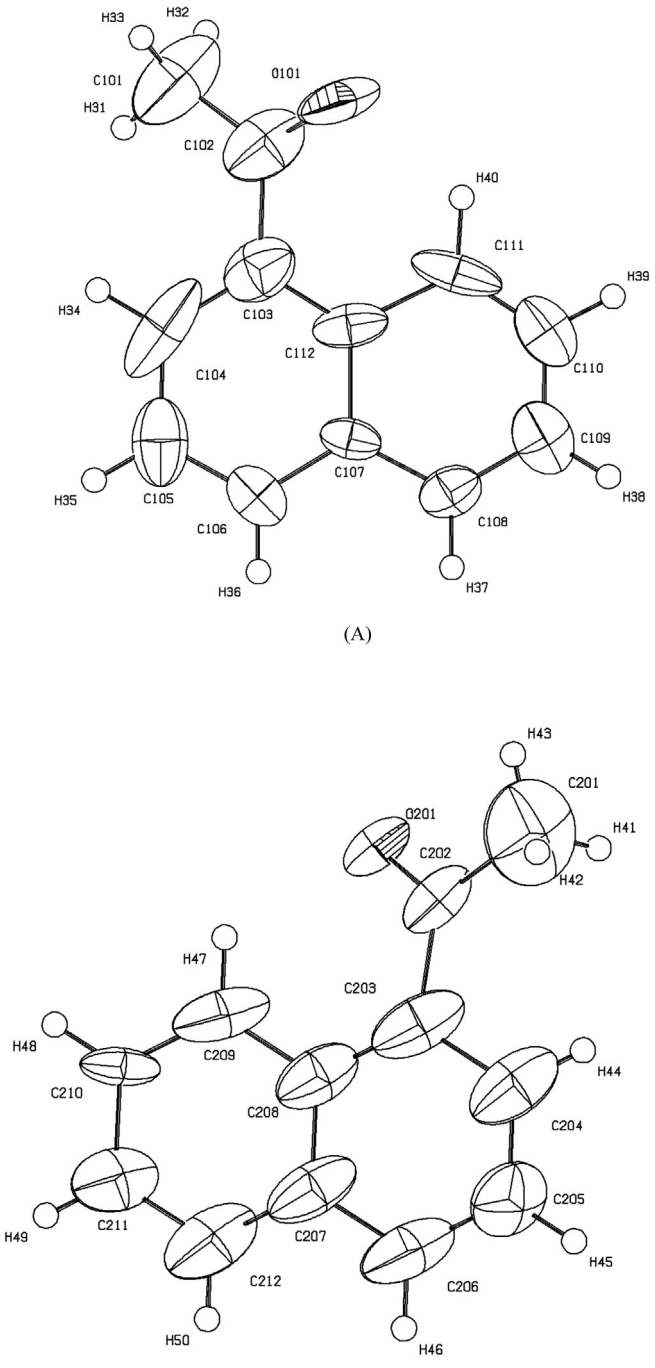

Fig. 5. Displacement ellipsoid plots showing the two crystallographically independent guest molecules in channel A & B at 93 K. Their site occupation numbers are 0.793(4) and 0.769(5) respectively. The ellipsoids are drawn at the 50% probability level. The bond distances and displacement parameters for the guest molecule in the B channel are heavily constrained to be similar and similar as in Ref. [17].

4.4. Evaluation

It might be difficult to achieve 100% removal of the original solvent from the target sites. As a result, refinement models that do not take their contribution into account will lead to systematic errors in the target molecule parameters. The use of refinement constraints & restraints to standard values may avoid poor geometry but make that geometry largely meaningless as independent information. Also the identification of a density peak as C, N or O, when not known for sure by other methods, as might be the case with natural products, can be challenging.

5. Valid versus value

A structure report based on poor but best attainable experimental data may still be valid as long as all experimental details and limitations are documented and commented upon. Its value lies in the validity for its intended use. The successful use of constraints and restraints to model disordered solvents may improve the value of the main part of the structure of interest. Contrarily, the need to use constraints and restraints on the geometry of the molecule of interest may severely lower its scientific value, in particular when the interest lies in geometrical details such as distances, angles and intermolecular interactions or the positive identification of unknown atom types from the geometry and peak density. The structure reported in [17] might be (made) valid and have a value in demonstrating the sponge technique but of very limited value for detailed geometry information.

6. So what makes a structure report valid?

Contrary to the early days of single crystal X-ray structure determination, many steps in the structure determination process are currently 'black boxed' and automated. This extends from proprietary data collection and data reduction software to refinement programs that are only available in executable form. The build-in assumptions, algorithms, limitations and pitfalls of the software tools that are used may not always be known to their casual user. A structure report should therefor not only include the author's interpretation of the experimental data but also all details of the analysis, including the primary diffraction data. All non-standard procedures, including the applied refinement constraints and restraints, should be detailed and all non-standard results reported and discussed. Only then, proper evaluation by referees, authors and users of the reported science will be meaningful and possible. The value of a valid report depends on its scientific usefulness. A heavily disordered and constrained structure may have limited value and is usually excluded from statistical studies with data archived in the CSD.

References

[1] C.R. Groom, I.J. Bruno, M.P. Lightfoot, S.C. Ward, Acta Cryst. B72 (2016) 171–179.
[2] S.R. Hall, F.H. Allen, I.D. Brown, Acta Cryst. A47 (1991) 655–685.
[3] S. Hall, B. McMahon, International Tables for Crystallography Volume G, 2005.
[4] A.L. Spek, J. Appl. Cryst. 36 (2003) 7–13.
[5] A.L. Spek, Acta Cryst. D65 (2009) 148–155.
[6] G.M. Sheldrick, Acta Cryst. C71 (2015) 3–8.
[7] L.J. Bourhis, O.V. Dolomanov, R.J. Gildea, J.A.K. Howard, H. Puschmann, Acta Cryst. A71 (2015) 59–75.
[8] L.M.J. Kroon-Batenburg, J.R. Helliwell, B. McMahon, T.C. Terwilliger, IUCrJ 4 (2017) 87–99.
[9] J.A. Labinger, Comptes Rendus Chemie 5 (2002) 235–244.
[10] A. Ienco, M. Caporali, F. Zanobini, C. Mealli, Inorg. Chem. 48 (2009) 3840–3847.
[11] D.A. Clemente, Inorg. Chim. Acta 358 (2005) 1725–1748.
[12] C.K. Johnson, ORTEPII, Report ORNL-5138. Oak Ridge National Laboratory, Tennessee, USA, 1976.
[13] M.A. Macias, L.M. Acosta, C.M. Sanabria, A. Palma, P. Roussel, G.H. Gauthier, L. Suescun, Acta Cryst. C71 (2015) 363–372.
[14] I.J. Bruno, J.C. Cole, M. Kessler, J. Luo, W.D.S. Motherwell, L.H. Purkis, B.R. Smith, R. Taylor, R.I. Cooper, S.E. Harris, A.G. Orpen, J. Chem. Inf. Comput. Sci. 44 (2004) 2133–2144.
[15] M.C. Etter, P.W. Baures, J. Am. Chem. Soc. (1988) 639–640.
[16] Y. Inokuma, S. Yoshioka, J. Ariyoshi, T. Arai, Y. Hitora, K. Takada, S. Matsunaga, K. Rissanen, M. Fujita, Nature 495 (2013) 461–466.
[17] Y. Inokuma, K. Matsumura, S. Yoshioka, M. Fujita, Chem. Asian J. 12 (2017) 208–211.
[18] A.L. Spek, Acta Cryst. C71 (2015) 9–18.