Improved crystal structure solution from powder diffraction data by the use of conformational information


It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1107/S1600576717012596

Publisher: Wiley

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur
CentAUR
Central Archive at the University of Reading
Reading's research outputs online
Improved crystal structure solution from powder diffraction data by the use of conformational information

Elena A. Kabova, Jason C. Cole, Oliver Korb, Adrian C. Williams and Kenneth Shankland

Improved crystal structure solution from powder diffraction data by the use of conformational information

Elena A. Kabova,* Jason C. Cole, Oliver Korb, Adrian C. Williams and Kenneth Shankland

*School of Pharmacy, University of Reading, Whiteknights Campus, Reading, Berkshire RG6 6AD, UK, bCambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK, and cSchool of Chemistry, Food and Pharmacy, University of Reading, Whiteknights, Reading, Berkshire RG6 6AP, UK. *Correspondence e-mail: e.kabova@reading.ac.uk

The effect of introducing conformational information to the DASH implementation of crystal structure determination from powder diffraction data is investigated using 51 crystal structures, with the aim of allowing increasingly complex crystal structures to be solved more easily. The findings confirm that conformational information derived from the Cambridge Structural Database is indeed of value, considerably increasing the chances of obtaining a successful structure determination. Its routine use is therefore encouraged.

1. Introduction

Global optimization (GO)-based methods of crystal structure determination from powder diffraction data (SDPD) make explicit use of a significant amount of chemical knowledge; well characterized bond lengths and bond angles are typically held as fixed values throughout the optimization. Furthermore, certain torsion angles (such as those spanning a double bond) and certain components of molecules (for example, cyclic groups) are often treated as fixed entities. Any remaining torsion angles around which atoms are free to rotate are treated as variables to be determined by the GO procedure and are allowed to vary freely in the range of 0–360°. Thus the conformational space of a molecule under study is treated as a continuum, rather than as a sequence of isolated conformations. This work sets out to improve the performance of SDPD by implementing conformational restraints or conformational bias, derived from observed crystal structures stored in the Cambridge Structural Database (CSD; Allen, 2002; Groom et al., 2016), to these freely varying torsion angles.

The utility of conformational information as constraints during crystal structure solution has long been recognized and has found particular application in macromolecular crystallography. For example, a protein molecule from a known crystal structure is often used as a start point for the crystal structure refinement of a distinct but closely related structure; see for example Scapin (2013) and DiMaio et al. (2011) and references therein. However, in the area of small-molecule crystallography, and in particular SDPD, conformational information has not been routinely employed, despite the fact that some work has demonstrated that it can be beneficial (CCDC, 2015; Florence et al., 2005; Middleton et al., 2002; Cole et al., 2014). Generally though, this evidence base is not strong, consisting of a few ‘one-off’ demonstrations and lacking.
Table 1
The 51 crystal structures used in this work.

<table>
<thead>
<tr>
<th>Code</th>
<th>Compound name</th>
<th>CSD refcode</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A7</td>
<td>Zopiclone</td>
<td>CUIHNEY10</td>
<td>Borea et al. (1987)</td>
</tr>
<tr>
<td>A16</td>
<td>Tolbutamide</td>
<td>ZZPFPUS02</td>
<td>Donaldson et al. (1981)</td>
</tr>
<tr>
<td>A18</td>
<td>Pigment orange 36 (PO 36)</td>
<td>HOYVOH</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>A20</td>
<td>Famotidine</td>
<td>FOGVIG03</td>
<td>Florence et al. (2003)</td>
</tr>
<tr>
<td>A21</td>
<td>Sotalol hydrochloride</td>
<td>SOTALC</td>
<td>Gadret et al. (1976)</td>
</tr>
<tr>
<td>A22</td>
<td>Glipizide</td>
<td>SAXFED</td>
<td>Burley (2005)</td>
</tr>
<tr>
<td>A23</td>
<td>Dilatazem hydrochloride</td>
<td>CEYHUZ01</td>
<td>Kojicprodic (1984)</td>
</tr>
<tr>
<td>A24</td>
<td>Zopiclone dihydrate</td>
<td>UCUVET</td>
<td>Shankland et al. (2001)</td>
</tr>
<tr>
<td>A25</td>
<td>Capsaicin</td>
<td>FABVAFO1</td>
<td>David et al. (1998)</td>
</tr>
<tr>
<td>A26</td>
<td>Pigment yellow (PY 181 polymorph β)</td>
<td>GITWUC</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>A28</td>
<td>Sodium 4-[(E)-4-hydroxyphenylidiazeny] benzene sulfonate dihydrate</td>
<td>YAYWUQ</td>
<td>Kennedy et al. (2001)</td>
</tr>
<tr>
<td>A30</td>
<td>Carbamazepine:indomethacin 1.1</td>
<td>LEZKOMZ01</td>
<td>Majumdar et al. (1985)</td>
</tr>
<tr>
<td>A31</td>
<td>2-[3-(2-Phenyloxethoxy)propyl sulfonyl] ethyl benzoate</td>
<td>BIFERO</td>
<td>Florence et al. (2005)</td>
</tr>
<tr>
<td>A32</td>
<td>S-Ibuprofen</td>
<td>JENKNOC10</td>
<td>Freer et al. (1993)</td>
</tr>
<tr>
<td>A33</td>
<td>Ampicillin trihydrate</td>
<td>AMPCHI01</td>
<td>Burley et al. (2006)</td>
</tr>
<tr>
<td>A34</td>
<td>Verapamil hydrochloride</td>
<td>CURHOM</td>
<td>Carpy et al. (1985)</td>
</tr>
<tr>
<td>A35</td>
<td>Amodiaquinum dichloride dihydrate</td>
<td>SENJIF</td>
<td>Linnás et al. (2006)</td>
</tr>
<tr>
<td>A36</td>
<td>Nifedipine (polymorph C)</td>
<td>BICCIZ01</td>
<td>Bortolotti et al. (2011)</td>
</tr>
<tr>
<td>A37</td>
<td>N-[2-(4-Hydroxy-2-oxo-2,3-dihydro-1,3-benzothiazol-7-yl)ethyl]-3-[2-(2-naphthalen-1- ylthoxy)ethylsulfonyl]propylaminium benzoate</td>
<td>PAHFI0</td>
<td>Johnston et al. (2004)</td>
</tr>
<tr>
<td>A39</td>
<td>Cyheptamide</td>
<td>TEVSOD01</td>
<td>Florence et al. (2008)</td>
</tr>
<tr>
<td>A40</td>
<td>Ornidazole</td>
<td>NETRUZ</td>
<td>Shin et al. (1995)</td>
</tr>
<tr>
<td>B21</td>
<td>Bis-[4-[2-(p-tosylamino)benzylideneamino]-2,3-benzo-15-crown-5,N,N',O]-copper(II)</td>
<td>RIFVAM</td>
<td>Dorokhov et al. (2007)</td>
</tr>
<tr>
<td>B27</td>
<td>4-(Phenylidiazenyl)napthalen-1-amine hydrochloride</td>
<td>OICANAN</td>
<td>Yatsenko et al. (2001)</td>
</tr>
<tr>
<td>B31</td>
<td>Telmasartan (polymorph A)</td>
<td>XUHHO01</td>
<td>Dinnabier et al. (2000)</td>
</tr>
<tr>
<td>B34</td>
<td>Clarithromycin (polymorph I)</td>
<td>NAVSUY02</td>
<td>Noguchi et al. (2012)</td>
</tr>
<tr>
<td>B35</td>
<td>Pigment orange 62(PO 62)</td>
<td>HOYVUPO</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>B36</td>
<td>Pigment yellow (PY 151)</td>
<td>HOYWAUPO</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>B37</td>
<td>Pigment yellow (PY 154 polymorph α)</td>
<td>HOYWEYPO</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>B38</td>
<td>Pigment yellow 194 (PY194)</td>
<td>HOYVICPO</td>
<td>van de Streek et al. (2009)</td>
</tr>
<tr>
<td>B39</td>
<td>2,4-Dinitro-N-phenyl-6-(phenylazo)benzamide</td>
<td>IHESUPIO</td>
<td>Chernyshev et al. (2002)</td>
</tr>
<tr>
<td>B40</td>
<td>N-Methyl-2,4-dinitro-N-phenyl-6-(phenylazo)benzamide</td>
<td>IHETEUP</td>
<td>Chernyshev et al. (2002)</td>
</tr>
<tr>
<td>B42</td>
<td>Trihexyphenidyl hydrochloride</td>
<td>KUZDITPO</td>
<td>Maccaroni et al. (2010)</td>
</tr>
<tr>
<td>B43</td>
<td>N-(2-Methoxyphenyl)-2-(2-methoxyphenylazo)-4,6-dinitrobenzamide</td>
<td>IHETAPO</td>
<td>Chernyshev et al. (2002)</td>
</tr>
<tr>
<td>B44</td>
<td>Nimustine hydrochloride</td>
<td>WAWZAXPO</td>
<td>Bekó et al. (2012)</td>
</tr>
<tr>
<td>B45</td>
<td>(R)-1-Phenylylammonium (R)-2-phenylbutyrate (polymorph II)</td>
<td>PBUPEAO1</td>
<td>Fernandes et al. (2007a)</td>
</tr>
<tr>
<td>B46</td>
<td>(R)-1-Phenylylammonium (R)-2-phenylbutyrate (polymorph III)</td>
<td>PBUPEAO2</td>
<td>Fernandes et al. (2007b)</td>
</tr>
<tr>
<td>B47</td>
<td>Tetracaine hydrochloride</td>
<td>XISVOKPO</td>
<td>Nowell et al. (2002)</td>
</tr>
<tr>
<td>B48</td>
<td>α/β-Lactose</td>
<td>LAKKEOPO</td>
<td>Lefebvre et al. (2005)</td>
</tr>
<tr>
<td>B49</td>
<td>N-(6-Phenylhexanoyl)glycyltryptophanamide</td>
<td>FEFFNOVPO</td>
<td>Bushmarinov et al. (2012)</td>
</tr>
<tr>
<td>B50</td>
<td>Pigment yellow 183 (PY183 polymorph α)</td>
<td>HOMMECO1</td>
<td>Ivashkevskaya et al. (2009a)</td>
</tr>
<tr>
<td>B51</td>
<td>Pigment yellow 191 (PY191 polymorph α)</td>
<td>HOMMIC01</td>
<td>Ivashkevskaya et al. (2009a)</td>
</tr>
<tr>
<td>B52</td>
<td>Pigment yellow 191 (PY191 polymorph β)</td>
<td>HOMMOM01</td>
<td>Ivashkevskaya et al. (2009a)</td>
</tr>
<tr>
<td>B53</td>
<td>Lisinopril dihydrate</td>
<td>GERWUX01</td>
<td>Sorrenti et al. (2013)</td>
</tr>
<tr>
<td>B54</td>
<td>Prednisolone succinate</td>
<td>KIXDEB01</td>
<td>Nishibori et al. (2008)</td>
</tr>
<tr>
<td>B55</td>
<td>Cytename (polymorph II)</td>
<td>SODNOPER</td>
<td>Florence et al. (2008)</td>
</tr>
<tr>
<td>B56</td>
<td>Carvedilol dihydrogen phosphate propan-2-ol solvate</td>
<td>PUTJOEPO</td>
<td>Chernyshev et al. (2010)</td>
</tr>
<tr>
<td>B57</td>
<td>Ritonavir</td>
<td>YIOPIO01</td>
<td>Bauer et al. (2001)</td>
</tr>
<tr>
<td>B59</td>
<td>n-Sorbitol</td>
<td>GLICD03P</td>
<td>Chernyshev et al. (2011)</td>
</tr>
<tr>
<td>B60</td>
<td>Chlorothiazide N,N-dimethylformamide solvate</td>
<td>NILESEP0</td>
<td>Fernandes et al. (2007)</td>
</tr>
<tr>
<td>B61</td>
<td>1,2,3-Tris(nondecanoyl)glycerol (polymorph β)</td>
<td>MEZNAPO</td>
<td>Helmholdt et al. (2002)</td>
</tr>
</tbody>
</table>

potential of exploiting conformational information in a more systematic and wide-ranging study, to provide a definitive report of its benefits.

2. Materials and methods

2.1. Selection and composition of powder X-ray diffraction data sets

A recent publication (Kabova et al., 2017) described the optimization of the key simulated annealing (SA) parameters in DASH (David et al., 2006) using 101 crystal structures. All those structures for which a success rate of less than 60% was achieved by SDPD, it is therefore timely to re-visit the

In the sense that there is no need to perform additional practical experiments, e.g. solid-state NMR.

A recent publication (Kabova et al., 2017) described the optimization of the key simulated annealing (SA) parameters in DASH (David et al., 2006) using 101 crystal structures. All those structures for which a success rate of less than 60% was
obtained in that work [using the default DASH SA parameters of cooling rate (CR) = 0.02, $N_1 = 20$ and $N_2 = 25$] were selected for subsequent evaluation in this current work. Details of these 51 structures are given in Table 1.

### 2.2. Software and hardware

The software and hardware employed in this work are summarized in Tables 2 and 3.

#### 2.2.1. Mogul

Mogul (Bruno et al., 2004) is a knowledge-based library of molecular geometries derived from the CSD. It acts as a source of information on preferred molecular geometries and can be used to validate the geometry of a solved structure. In the case of SDPD, our main interest is in obtaining information on preferred conformations of a molecule under study, based on the torsion angle distribution information contained in Mogul. Taking the C6—C5—O2—C20 torsion angle of verapamil hydrochloride (structure A34) as an example, the distribution of structurally closely related torsion angles (based on circa 11 100 CSD-deposited crystal structures) is shown in Fig. 1. The distribution clearly shows that this torsion angle is likely to adopt a value in the range $-20^\circ$ to $+20^\circ$, as approximately 95% of deposited crystal structures with this structural feature fall into this range.

#### 2.2.2. Use of Mogul distributions as hard constraints within DASH

During the GO process, any torsion angle that is free to rotate in the molecule under study can be subjected to a Mogul query by pressing the ‘Modal’ button on the ‘parameter bounds’ window of DASH. From the results of the Mogul query, a set of discrete constraints is derived. For example, in the case of the C6—C5—O2—C20 torsion angle of verapamil hydrochloride, only values in the ranges of $0^\circ$ to $+20^\circ$ and $-20^\circ$ to $0^\circ$ are permitted (Fig. 3). The reduction in search space from 360$^\circ$ to only 40$^\circ$ for one torsion angle is not expected to have a notable impact on the overall success rate.
in solving verapamil hydrochloride, but if similar Mogul-derived restrictions are applied to all of the 14 variable torsion angles in the molecule, then the total search space reduction becomes more significant.

2.2.3. Use of Mogul distributions to bias parameter space sampling within DASH. Mogul distribution bias (MDB) is an alternative method of exploiting the Mogul-derived conformational information. In contrast to the Mogul-derived constraint approach, the MDB approach still samples the full \(0–360\) range for a torsion angle according to a Maxwellian distribution. However, here the Maxwellian is binned and multiplied by the value of the corresponding bin in the Mogul distribution, to generate a new distribution which favours moves to torsion angles in regions of space that are heavily populated in the Mogul distribution. MDB is invoked within DASH by pressing the ‘Set MDB’ button in the ‘parameter bounds’ window (Fig. 2). This automatically performs the necessary Mogul searches on all torsion angles that are free to rotate in the molecule under study. Considering again the \(C6–C5–O2–C20\) torsion angle of verapamil hydrochloride, the Mogul search returns a string [hidden to the interactive user, but visible in the DASH batch file (DBF) that is generated for a batch run]: 4.6355 MDB -180 180 18 8072 2245 446 113 34 18 9 10 10 9 15 18 28 15 14 16 14 14, where the initial torsion angle value\(^3\) is listed first, followed by the instruction to use the MDB approach, the minimum and maximum angular values, the number of bins in the probability histogram, and finally the number of observations in each bin.

The MDB and Mogul hard constraints approaches introduce the same underlying information in different ways, resulting in a different exploration of \(\chi^2\) search space during the SA.

2.3. Performance analysis

An empirical log-of-the-odds (ELO) analysis was performed in order to evaluate any increase in the success rate (SR) as a result of the conformational information introduced by the use of Mogul.

The ELO, described by Cox & Snell (1989), can be written as

\[
ELO = \ln \left( \frac{\text{SR} + 0.5}{100 - \text{SR} + 0.5} \right)
\]

Logistic regression was performed using Minitab. Full details are given by Kabova et al. (2017).

3. Experimental

Following the protocol established previously, 50 SA runs were initially executed on all structures, using optimized SA control parameters of \(CR = 0.27, N_1 = 73\) and \(N_2 = 56\) (Kabova et al., 2017). Each run was set to perform \(1 \times 10^7\) SA moves followed by a short simplex calculation. A \(\chi^2\) multiplier of 1 ensured the full number of SA steps was always carried out and the SA was not prematurely terminated. Five hundred SA runs of \(5 \times 10^7\) moves were performed for compounds for which no successful solution was observed in the initial 100 SA runs. The MDAH utility (Griffin et al., 2009) was used to distribute these longer runs over ten CPU cores. The option to manually alter any of the torsion angle ranges suggested by Mogul was not used. For consistency and to facilitate comparison of success rates, the same values of the random seeds in DASH were used throughout.
The overall benefit of using MDB or Mogul constraints during the SA process is most clearly shown by the ELO analysis (Fig. 4) performed on all the results obtained in this work. The shift to the right of the fitted curves, relative to the conventional DASH approach using optimized SA parameters, shows the increased probability of solving a crystal structure owing to the inclusion of Mogul-derived information. These gains, which were obtained in combination with DASH’s optimized SA parameters (Kabova et al., 2017), were also realized when DASH’s default SA parameters were used [results not shown here; see Kabova (2016) for full details].

Interestingly, the use of torsion angle constraints which did not span the torsion angle values seen in the final crystal structure did not necessarily preclude obtaining a good solution; the simplex minimization employed at the end of the SA ignores the Mogul constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.

Because of the chemical diversity of structures in the CSD, and the numerous factors that influence their packing into a crystal structure, it is inevitable that novel crystal structures will possess torsion angles that are not well represented in the CSD. Even though there are nearly 900 000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so either their MBD influence on the SA is minimal or no valid constraints and gives the possibility that the correct torsion angle values can be recovered. For example, with structure A36, five of the 12 torsion angle constraints did not span the crystal structure values, but the mean absolute error remained the same. An increase in the SR with both constraints and MBD was still seen. Unsurprisingly, the MBD approach deals better with such cases than the constraint-based approach; an MBD distribution does not explicitly preclude a parameter taking an unlikely value during the SA itself, it merely makes it less likely that it is sampled.
Such cases (around 10% of the total torsion angles of this work) are treated as fully flexible by DASH. In the case of B57 [ritonavir form II, 28 degrees of freedom (DoF)], the complex molecule was reported to have an unexpected conformation in the crystal structure, as a result of a strong hydrogen-bonding network (Bauer et al., 2001). Whilst the use of conformational information did not allow DASH to solve the structure with the use of its default SA settings, a solution was obtained with the optimized SA parameters using both MBD and Mogul-derived constraints, although the latter required the use of 500 SA runs.

When using Mogul-derived torsion angle information, it is important to consider the way in which torsion angles are defined in the input model Z-matrices. If, for example, the input Z-matrix contains a torsion angle that is defined using at least one hydrogen atom, then no Mogul distribution is generated and potential information is lost. Taking B56 (a structure for which no solution was obtained in the absence of torsion angle information) as an example, five of the ten torsion angles in the Z-matrix (as automatically generated by DASH) are described with the use of a hydrogen atom and as such are ineligible for inclusion in the Mogul/MBD distributions. This represents a considerable loss of information, but despite this, two correct solutions were found using MBD with 500 SA runs. A simple workaround for this hydrogen-related issue, when using a CIF input model, is to manually re-order atoms in the CIF, such that all the hydrogen atoms appear at the end of the atom list. Future releases of DASH may address this issue by changing its Z-matrix generating code, or by giving the option to include filtered results in a distribution.

The reduction in the SR observed for a small number of compounds (e.g. A37, A40 and B60) must be addressed. In the cases of A40 and B60, the DoF in the problem are largely positional and orientational (18 out of 30 for A40, 36 out of 42 for B60), and as such they are not so heavily influenced by the introduction of conformational information. Interestingly, even when the correct conformations of the three independent molecules of A40 are used as input, and held fixed throughout the SA, DASH fails (with the default settings) to solve the structure within 50 SA runs, indicating the extent of the positional/orientational challenge for this structure.

6. Conclusions

This work represents a comprehensive study of the effects of including conformational information, derived from the CSD, on SDPD using the DASH program. The results provide strong evidence that such information should be routinely employed when faced with complex structures; the necessary tools are already in place (the fully automated MBD option is particularly convenient) and there is no significant computational overhead involved in its use. It is likely that other GO-based approaches to SDPD can benefit from this type of information and the tools provided in the CSDS are extremely valuable in this regard.

7. Availability and documentation

Details of DASH’s availability can be found at https://www.ccdc.cam.ac.uk/solutions/csd-materials/components/dash/.

Acknowledgements

EAK thanks the University of Reading and the Cambridge Crystallographic Data Centre (CCDC) for funding. We thank Mark Spillman and David Edgeley for their help with various computational matters pertaining to the rapid execution of DASH, and Wei Dong for his help in implementing MBD within DASH. We are also grateful to the University of Reading Chemical Analysis Facility for local powder diffraction facilities.

References


*DASH uses a Z-matrix description (Shankland, 2005) of the molecule internally, with flexible torsion angles flagged as variables for optimization by the SA. If a model is input to DASH as a CIF, MOL, MOL2, RES or PDB, it is automatically converted to a Z-matrix.

The angle is ‘filtered’, on the basis that H-atom positions are often fixed geometrically, are ‘assumed’ or are otherwise unreliable.