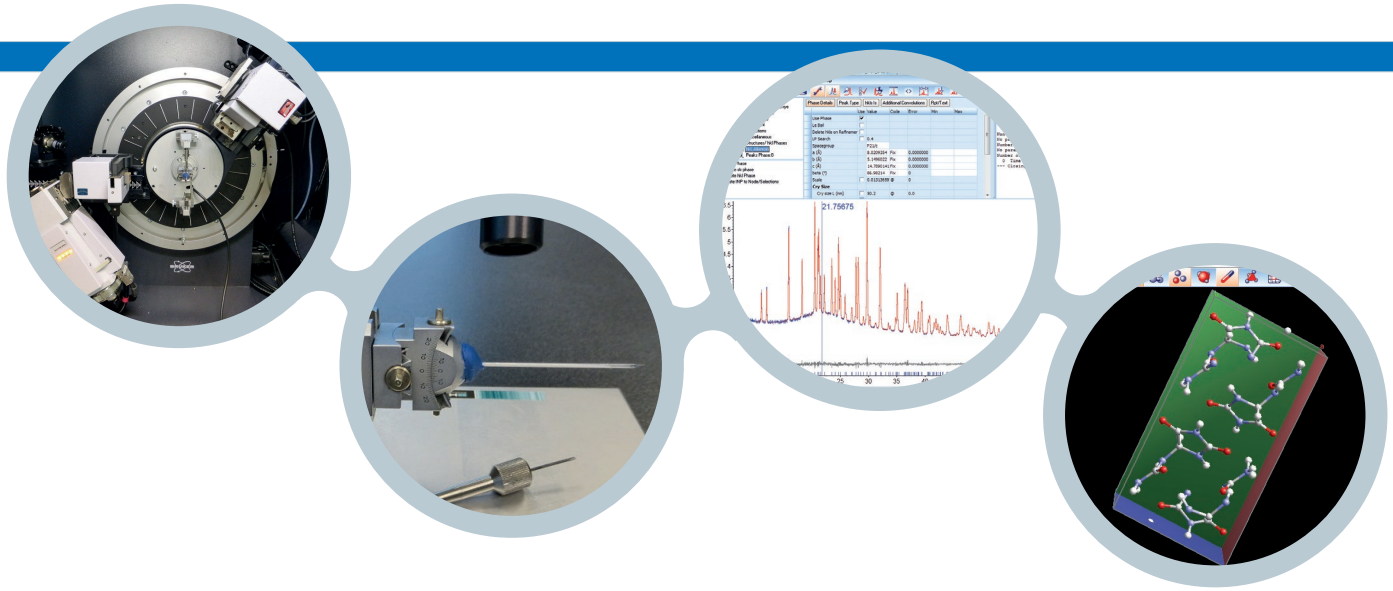

Application Note XRD 604

# Structure Solution from X-ray Powder Diffraction Data for Pharmaceutical Samples - a Walk-Through Example using a Rigid Body for the Molecular Structure

## Introduction

Crystal structure solution for pharmaceutical compounds is not always possible by traditional single crystal X-ray diffraction techniques. This occurs, in particular, when metastable polymorphic forms can only be obtained as micro-crystalline aggregates or when crystals exhibit excessive twinning that can lead to severe problems with single crystal diffraction. In some cases compounds may only be available as powders or non-ambient studies are only feasible using powders leaving Structure Determination from Powder Data (SDPD) as the only viable option. While this method suffers from the loss of information due to the collapse of three-dimensional diffraction data to one-dimensional powder data, it has the advantage of being more representative of the bulk sample. It will typically only succeed for single phase samples without any unknown impurity phases.

This application note describes how to collect and interpret good quality powder diffraction data for a pharmaceutical sample. It provides a guide line for data collection as well as structural modelling using the DIFFRAC.TOPAS software. While many pharmaceutical laboratories have the necessary hardware and software, relatively few industrial users seem to be successful with the SDPD method. It is often perceived as being complicated, in part because it also involves using the Launch mode of DIFFRAC.TOPAS, which is based on more flexible text input files. For that reason the process is described for a small molecule compound in a step by step manner from data collection to the various steps of data processing. Ideally, this can serve as a template and avoid some of potential frustration with syntax errors in input files.

The compound Allantoin was chosen as an example for a typical small molecule where the molecular structure is already known and can be entered as a "rigid body". The chemical formula for Allantoin is $C_4H_6N_4O_3$ and the molecular structure can be downloaded from various sources, which makes it a convenient test sample.

## Instrumental Setup and Data Collection

For most pharmaceutical samples, preferred orientation effects prevent the measurement of accurate relative peak intensities in Bragg-Brentano geometry. That typically means that data should be collected in a configuration where the powder sample is filled into a capillary that rotates during data collection to randomize crystallite orientations. In this example, a D8 ADVANCE system equipped with a copper X-ray tube, LYNXEYE XE detector and capillary stage was used. The incident beam optic could be a Johansson primary beam monochromator or a Goebel mirror (focussing or parallel). In this report the data was collected using a primary beam monochromator with an approximate focusing length of 360 mm. That means that the beam is several mm wide at the instrument center irradiating the whole capillary and focusing on the detector. Large 1 mm diameter capillaries can be used for low density pharmaceutical samples leading to higher peak intensities. The LYNXEYE XE detector was close enough to the capillary to cover approximately 4.5° with the detector opening. Two air-scatter screens should be used with the capillary stage to minimize the background from air-scatter at low angles. They can be mounted off center from the capillary to avoid touching the capillary during mounting. The lower beam knife should be adjusted to block the direct beam. If the upper beam knife is mounted off center, the scan type "2Theta scan" has to be selected with the tube parked at 0°. This works well for a 2Theta range up to approx. 70° 2Theta. For measurements to higher scan angles "coupled 2Theta/Theta" scans are necessary. In that case, the spacer in the upper beam knife has to be removed and the knife has to be mounted centered just above the capillary to avoid blocking the incident beam (see Figure.1)

For structure solution or refinements it is generally an advantage to collect data in a wide 2Theta range to get as much information as possible from the sample. Unfortunately, the intensity drops off considerably to higher angles mostly due to the Lorentz Polarization (LP) factor and thermal vibrations leading to worse counting statistics in the high angle range. To collect data at higher angles with the same counting statistics, a variable count time (VCT) collection scheme can be used, which is implemented in the Wizard plugin of the measurement software (see Figure 2 and Figure 3). In this example, the count rate ramps from 1sec/step at low angles up to 1 sec/step at higher angles. The DIFFRAC.TOPAS software supports the VCT mode and scales the scan ranges appropriately with the y-axis shown in cps (per detector channel).

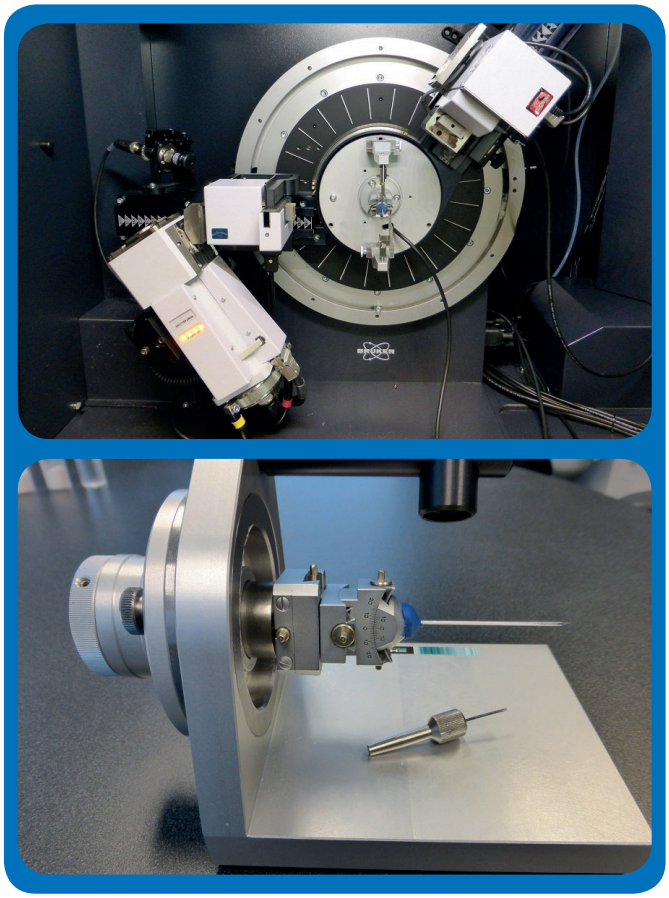

Figure 1 D8 ADVANCE configuration used for data collection with primary beam monochromator, capillary stage and LYNXEYE XE detector. Low absorption glass capillaries with 1 mm diameter were used, filled with Allantoin and sealed off with a lighter. The capillary was mounted on the goniometer head using modelling clay for support and centered in the beam using the alignment microscope to avoid any wobbling.

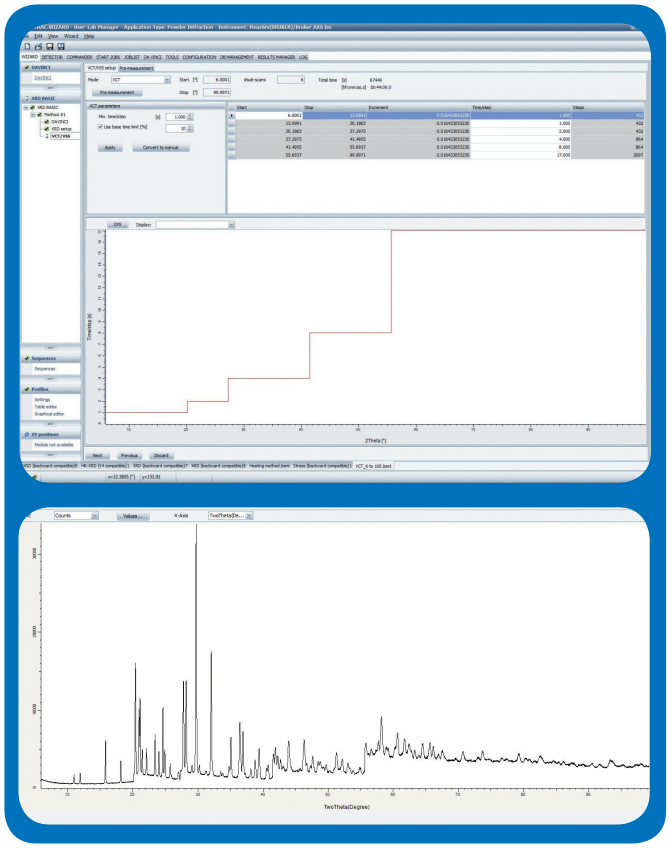

Figure 2 Variable Count time scheme, as defined in the Wizard plugin of the measurement software and corresponding data set shown in the results manager. The resulting steps in the data are automatically scaled in the DIFFREC.EVA or DIFFRAC.TOPAS software packages.

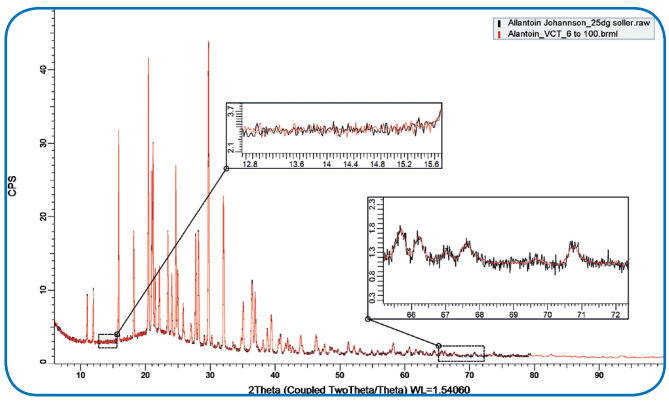


Figure 3 Comparison of scans with constant time/step (black) and variable count time (red) scheme from Figure 2. The improvement in counting statistics is evident at higher angles.

## Indexing the Diffraction pattern

The process of determining an unknown unit cell is usually referred to as Indexing and is the first step in the structure solution process. For the Least Square Indexing (LSI) Algorithm in DIFFRAC.TOPAS at least the first 20 peak positions have to be determined as accurately as possible. Missing peaks, impurity peaks in the pattern or excessive zero errors will drastically reduce the chances of finding the correct unit cell. If the data quality is sufficient, it is normally possible to find the correct cell volume and crystal system and in most cases also the correct space group.

To determine the peak positions accurately using profile fitting it is advantageous to constrain the peak width to the same value to be able to identify overlapping peaks easier. In this example the peaks were fitted using a Pseudo Voigt peak type with constrained peak widths and refined simple axial model to account for the peak asymmetry caused by axial divergence (See Figure 4).

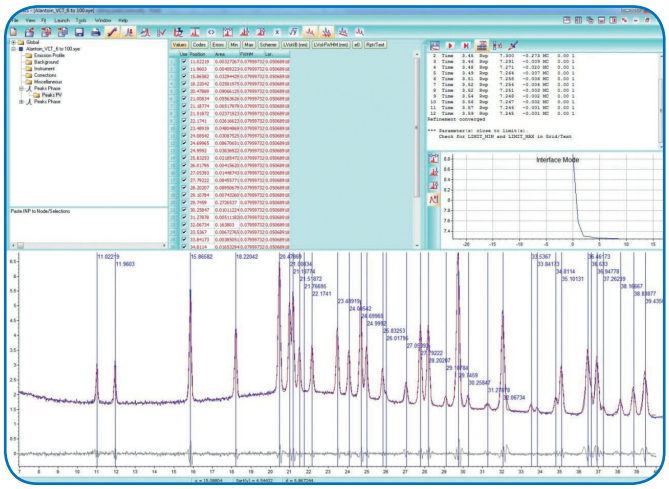

Figure 4 Peak Profile Fit in DIFFRAC.TOPAS for the first 30 peaks. The peak shapes were constrained to have the same FWHM and Lorentzian component parameter, Lor.

There are only minor peak overlaps in this example and only one additional peak was detected in the area around 35°. Once the peak profile fit is satisfactory, an Indexing range can be created directly from that peak list and the LSI algorithm can be started. It is useful to initially exclude the triclinic system from the list of possible crystal systems since it tends to obscure higher symmetry solutions. The indexing results are displayed in a table of possible unit cell solutions and a plot of the "goodness of fit" vs unit cell volume is shown below. Note, that the top entry with the highest GOF value will not always be the correct solution. Generally, it should be possible to find a solution that explains all the observed peaks; however, the best solution is the smallest unit cell that explains all measured peaks without any calculated additional peaks, which appear as dotted lines in the main display plot (Figure 5).

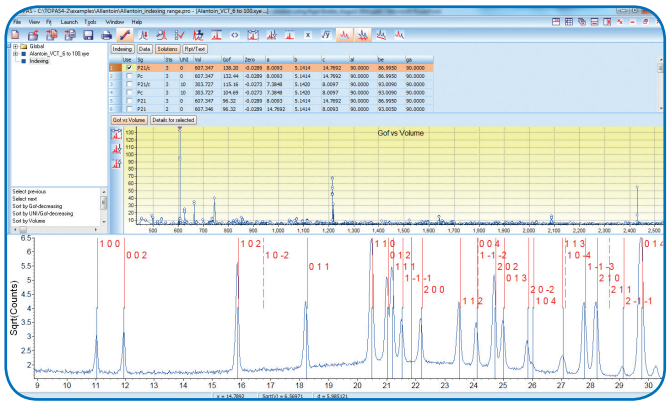


Figure 5 Results of the LSI Indexing algorithm using DIFFRAC.TOPAS. The solutions tab displays a list of possible unit cells and a graphical representation of the Goodness of Fit (GOF) versus the cell volume. The smallest unit cell that explains all observed peaks without any additional calculated peak positions (displayed as dashed lines) is likely the correct solution.

## Whole Powder Pattern Fitting and preparation of an Input file for Structure solution

After selecting the best indexing solution, a Pawley or LeBail fit can be used to obtain the best possible fit over the whole angular range. This step insures that the best lattice, peak shape and background parameters are determined before introducing any atomic sites. These parameters will not be refined in the subsequent structure solution run. The same Pseudo Voight peak type with refined simple axial divergence model parameter was used, but lattice parameters and sample displacement were refined. The refined background function should be checked visually to see if its shape looks reasonable, especially at higher angles where the large number of overlapping peak intensities may be correlated with the background polynomial. Once an optimum fit for the whole data set is obtained, the corresponding peak shape parameters, background parameters, and lattice parameters can all be fixed and the associated residual Rwp is the best possible residual that can be expected in any structure solution run. The Pawley fit is also the last step that can be performed in the Interface mode of DIFFRAC.TOPAS. The corresponding project file with all parameters fixed can be exported to a text file and becomes the starting point for a structure solution input file in the Launch mode of DIFFRAC.TOPAS (Figure 6) that can be edited in any text editor.

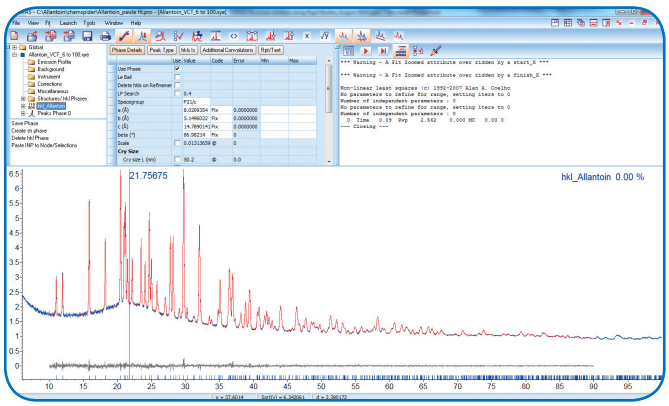


Figure 6 Pawley Fit of Allantoin. The amorphous halo from the glass capillary was fitted with a single broad peak. After checking for a sensible background shape, especially at high angles, all refinable parameters were fixed and the resulting project exported to an INP file, which serves as the starting point for the structure solution in the Launch mode of DIFFRAC.TOPAS.

For the structure solution run, no peak intensities will be refined directly anymore, but will be calculated from structural model parameters that are being refined. The following editing steps to the Input file are necessary to convert the Pawley Input file to an Input file for the structure solution attempt using atomic sites for each atom. The atomic positons will be constrained using a "rigid body" model for the molecule, because the molecular structure is already known.

1. Delete the hkl_phase intensities in the Pawley fit input file, essentially all lines in the text file that start with "hkl_m_d_th2".

2. Insert the "Auto_T(10)" macro on top of the text file. This macro contains simulated annealing and randomizing commands that will work in a wide range of structure solution applications. If no solution can be found, the temperature regime parameter can be changed.

3. Optionally, insert the "Decompose (0.005)" command, which will speed up the calculation by displaying data points at peak positions only.

4. Insert the "Str" keyword in front of the lattice and peak shape parameters that were determined during the Pawley fit. This keyword defines a structure in DIFFRAC.TOPAS.

5. Insert "scale @ 0" after the Str keyword to scale the calculated peak intensities. The @ symbol in an input file, with or without parameter name, indicates that this parameter will be refined.

6. Insert all 17 atomic sites below the structure in the following format, set all x y z coordinates to 0, all occupation factor to 1 and all isotropic temperature factors to 1.

| Key | Site |
|---|---|
| Label | O1 |
| Coordinates | x 0 |
| | y 0 |
| | z 0 |
| Key | occ |
| Atom type | O |
| Occ. Factor | 1 |
| Key | Beq |
| Temp. Factor | 1 |

7. Constructing a Rigid body for the Allantoin Molecule in Z-matrix notation. DIFFRAC.TOPAS allows the input of rigid bodies either as xyz-coordinates in a "point_for_site" notation or in a simple z-matrix notation, where each atom is introduced with distances and angles in relation to the neighboring atom using internal coordinates. For example, introducing six carbon atoms with interatomic distances of 1.3Å, bond angles of 120° and torsion angles of 0° will create a benzene ring using the following notation

```
rigid
z_matrix C1
z_matrix C2     C1   1.3
z_matrix C3     C2   1.3   C1   120
z_matrix C4     C3   1.3   C2   120   C1   0
z_matrix C5     C4   1.3   C3   120   C2   0
z_matrix C6     C5   1.3   C4   120   C2   0
```

Every numerical value in this notation can be parameterized, which makes the Z-matrix very practical, but it can be tedious to build up a larger molecule atom by atom. Typically, a molecule may be available in a variety of other molecular formats and free conversion programs such as "OpelBabel" can be used to convert to a format with x,y,z coordinates for each atom. In this example the Allantoin molecule was downloaded from the chemspider.com website in .mol format and converted with the "OpenBabel" program to a format with XYZ coordinates. By pasting the "Point_for Site" command in front of each atom, these coordinates can then be pasted directly into the rigid body editor of DIFFRAC.TOPAS. This will allow an automatic conversion into the Z-matrix format of DIFFRAC.TOPAS which can be copied and pasted directly into the input file (Figure 7).

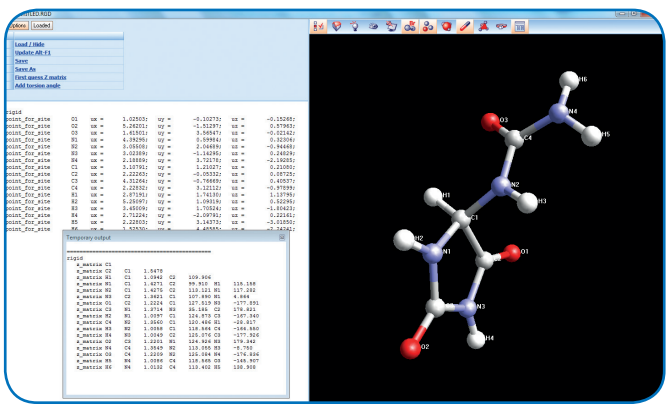

Figure 7 Rigid body editor in DIFFRAC.TOPAS. The point_for_site notation is useful to enter xyz coordinates. The "first Guess Z-matrix" option will convert to a z-matrix notation that can be parameterized and refined to make the model structure less rigid.

8. Add "Rotate_about_axies(@ 0, @ 0, @ 0)" and "Translate(@ 0, @ 0, @ 0)" commands to the input file to be able to move the molecule within the unit cell. All x, y and z coordinates are being refined.

9. Optionally, add "view_structure" to display the structure viewer. At this point the input file with a section displayed in Figure 8 is ready to be launched for structure solution. The only refinable parameters are the rotation and translation parameters for a completely rigid body and the resulting data fit and corresponding Rwp residual is poor and obviously far from correct.

```
      site O1   x  0 y 0  z 0   occ O  1   beq 1
      site O2   x  0 y 0  z 0   occ O  1   beq 1
      site O3   x  0 y 0  z 0   occ O  1   beq 1
      site N1   x  0 y 0  z 0   occ N  1   beq 1
      site N2   x  0 y 0  z 0   occ N  1   beq 1
      site N3   x  0 y 0  z 0   occ N  1   beq 1
      site N4   x  0 y 0  z 0   occ N  1   beq 1
      site C1   x  0 y 0  z 0   occ C  1   beq 1
      site C2   x  0 y 0  z 0   occ C  1   beq 1
      site C3   x  0 y 0  z 0   occ C  1   beq 1
      site C4   x  0 y 0  z 0   occ C  1   beq 1
      site H1   x  0 y 0  z 0   occ H  1   beq 1
      site H2   x  0 y 0  z 0   occ H  1   beq 1
      site H3   x  0 y 0  z 0   occ H  1   beq 1
      site H4   x  0 y 0  z 0   occ H  1   beq 1
      site H5   x  0 y 0  z 0   occ H  1   beq 1
      site H6   x  0 y 0  z 0   occ H  1   beq 1

rigid
   z_matrix C1
   z_matrix C2    C1    1.5478
   z_matrix H1    C1    1.0942  C2    109.906
   z_matrix N1    C1    1.4271  C2    99.910  H1    115.158
   z_matrix N2    C1    1.4275  C2    113.121 N1    117.282
   z_matrix N3    C2    1.3621  C1    107.890 N1    4.864
   z_matrix O1    C2    1.2224  C1    127.519 N3    -177.891
   z_matrix C3    N1    1.3714  N3    35.185  C2    178.821
   z_matrix H2    N1    1.0097  C1    124.873 C3    -167.340
   z_matrix C4    N2    1.3560  C1    120.486 H1    -28.817
   z_matrix H3    N2    1.0058  C1    118.564 C4    -164.550
   z_matrix H4    N3    1.0049  C2    125.076 C3    -177.926
   z_matrix O2    C3    1.2201  N1    124.926 N3    179.342
   z_matrix N4    C4    1.3549  N2    113.055 H3    -8.750
   z_matrix O3    C4    1.2209  N2    125.084 N4    -176.836
   z_matrix H5    N4    1.0086  C4    118.565 O3    -145.907
   z_matrix H6    N4    1.0132  C4    113.402 H5    138.908

Rotate_about_axies(@ 0, @ 0, @ 0)
Translate(@  0, @  0, @ 0)

view_structure
```

Figure 8 Section of the Input file with atomic sites, rigid body definition and translation and rotation command.

10. Since the initial model is obviously too rigid, refinable parameters for the torsion angles can be added by first naming and initializing the parameter and then adding those to the rigid definition. This is best done in the rigid body editor (Figure 9) to be able to directly see which torsion angles are affected.

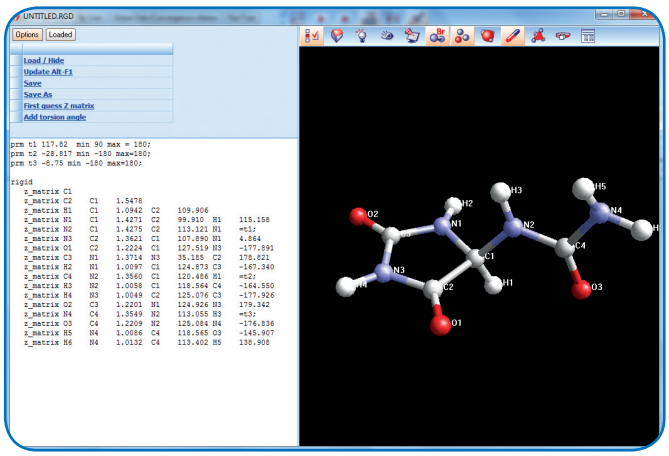


Figure 9 Rigid body definition with parameterized and refinable torsion angles

11. Run the modified input file and stop the refinement cycle once a few low Rwp values are found. The best solution will be displayed. At this point the only major misfit should the background at higher angles (Figure 10). This is not surprising for complex unit cells, where usually some correlation between the background function and numerous overlapping peaks at higher angles can be observed.

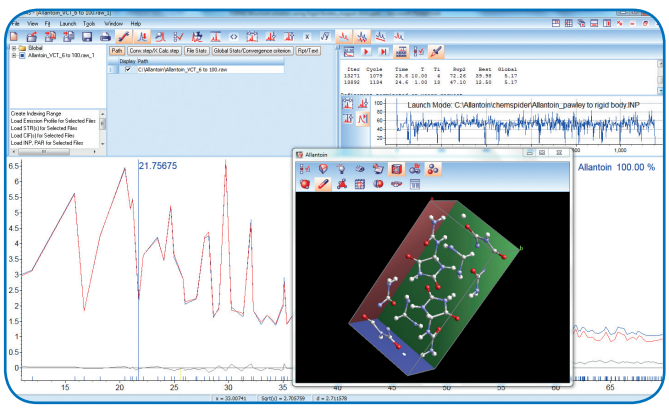


Figure 10 After refining three torsion angles the Rwp values drops significantly to approx. 12%. The only obvious misfit is the background at higher angles.

12. Comment out the “Auto_T” and “decompose” commands, fix the “Rotate_about_axies” and “Translate” commands by removing the @ symbols. Enter the @ symbol in front of the background polynomial function to refine the background function at higher angles. After running the refinement, the residual will drop to approximately Rwp = 4.4% (Figure11). This value is already close to the best residual obtained in the previous Pawley fit where all peak intensities were refined independently. That means that the obtained structure is not perfect yet but judging from the overall fit, it is likely good enough to be used in quantitative Rietveld analysis of mixtures. It can be further refined by adding more refinable parameters such as bond lengths or angles to the rigid body and as a last step even refinement of temperature factors. However, the risk of parameter correlation will increase with the number of parameters and ultimately only parameters that are supported by the data quality should be reported.

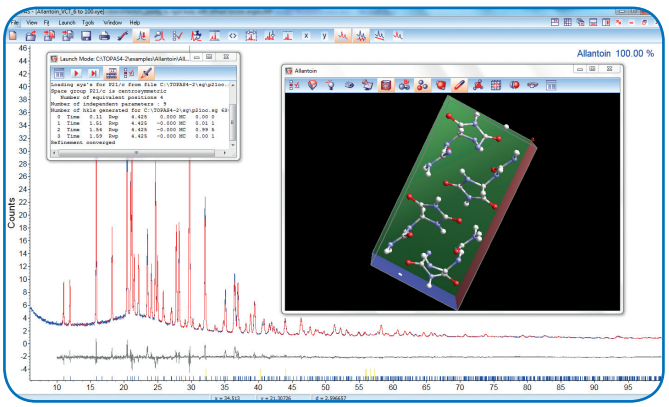


Figure 11 Result with Rwp =4.4% after refining only three torsion angles and the background function.

13. Adding the command Out_CIF_STR(Allantoin.CIF) to the bottom of the Input file will write a structure file in CIF format, that can be used in other programs.

## Conclusion

Solving crystal structures of pharmaceutical samples from good quality powder diffraction laboratory data is a worthwhile undertaking and can often be surprisingly straight forward, if the molecular structure is known and the data quality is sufficient. It generally only requires basic crystallographic knowledge to generate a structure file that is sufficiently accurate to be used for quantitative Rietveld analysis of phase mixtures. The resulting structure file may then be used for quantifying multiple polymorphs in a mixture or for quantification of active ingredients in pharmaceutical formulations.

**Bruker AXS GmbH**

Karlsruhe · Germany
Phone +49 721 50997-0
Fax +49 721 50997-5654
info.baxs@bruker.com

**www.bruker.com**