Enhanced structure determination from powder diffraction data via algorithm optimisation and the use of conformational information

Elena A. Kabova

Submitted for the Degree of

Doctor of Philosophy

School of Pharmacy
PO Box 226
Whiteknights
Reading
Berkshire
RG6 6AP

October 2015

Declaration

I confirm that this is my own work and all other materials from various sources are properly and fully acknowledged.

(Elena A. Kabova)

To my son Alexander, with love.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my PhD supervisors at the University of Reading (Dr Kenneth Shankland and Professor Adrian Williams) and at the CCDC (Dr Jason Cole and Dr Oliver Korb) for their invaluable input and support throughout the course of this work.

Financial support from the University of Reading and the CCDC is gratefully acknowledged.

Thanks are due to Dr Manuel López-Ibáñez for his support and advice with irace, to Dr Mark Spillman for the FDASH python script, to Mr Sandro Leidi for his advice on statistical matters and to Mr Nicholas Spencer for his encouragement and support with data collection.

I thank all my friends and colleagues in the Pharmacy Department at Reading University. My immediate team mates from the 'Shankland group', Dr Mark Spillman and Daniel Nicholls, are gratefully acknowledged for their direct and indirect contributions to my work. Special heartfelt thanks go to David Edgeley for *all* his help, ideas and brainstorming and - very importantly – all the doughnuts we shared.

Last but not least, I thank my family (immediate and extended) for their love and support throughout this project. Special thanks go to Carol Shipton for being an immense source of endless encouragement, motivation, inspiration and love.

Abstract

The performance of DASH has been evaluated against powder X-ray diffraction data collected from 101 molecular crystal structures, representing the most comprehensive testing of a "structure determination from powder diffraction data" (SDPD) program carried out to date. These 101 structures cover a broad range of molecular complexities, from very simple (6 degrees of freedom) to very challenging (49 degrees of freedom). 95 of the crystal structures could be solved with the current version of DASH, going some way to explaining why the parameterisation of its simulated annealing (SA) algorithm has not been altered since the launch of the program in 1999.

This thesis explores optimisation of key DASH SA parameters using the program irace. The irace runs, comprising 255,000 individual DASH runs and requiring approximately 1300 CPU days of compute time, produced six sets of SA parameters which differed greatly from the DASH default parameters and which markedly improved the performance of DASH. Further evaluation of these six sets against all 101 compounds (a further 2874 of days of CPU time), allowed selection of one best-performing set, which delivered an order of magnitude improvement in the success rate with which crystal structures were solved. The adoption of these parameter values as the defaults in future releases of DASH is strongly recommended and is expected to broaden the range of molecular complexities to which the program can be applied.

Three distinct approaches to further improving DASH performance, based on introducing prior conformational knowledge derived from the Cambridge Structural Database (CSD), have also been assesed. The findings show that inclusion of conformational knowledge brings significant additional gains in SDPD performance, and that existing implementations of these approaches in the DASH / CSD System are close to being ready for routine use.

Table of contents

Acknowledgements ..... **ii**

Abstract ..... **iii**

1 Introduction **1**

1.1 Solving crystal structures from powder diffraction data 3

1.1.1 The intrinsic problems of PXRD 3

1.1.2 The SDPD workflow 4

1.2 Methods for SDPD 5

1.2.1 Single-crystal structure solution methods adapted for powder diffraction data 6

1.2.1.1 Conventional and modified direct methods 6

1.2.1.2 Patterson methods 7

1.2.1.3 Charge flipping 7

1.2.1.4 Maximum entropy 8

1.2.2 Direct-space methods 9

1.2.2.1 Simulated annealing 10

1.2.2.2 Genetic algorithm 11

1.2.3 Other methods of SDPD 12

1.2.4 Summary of the SDPD methods 13

1.3 Powder diffraction and crystal structural complexity 13

1.4 SDPD and DASH 14

1.5 The current limits of PXRD 20

1.6 Enhancing SDPD 25

1.6.1 Computational gains 25

1.6.2 Parameter tuning 26

1.6.3 Prior conformational knowledge 26

1.7 Aims and objectives 27

1.7.1 Aims 27

1.7.2 Objectives 27

1.8 References 28

2 Materials and methods 35
2.1 Materials 36
2.1.1 Data sets for previously solved crystal structures 36
2.1.2 Laboratory X-ray data collection 36
2.2 Methods 46
2.2.1 Software 46
2.3 Hardware 46
2.4 References 48
3 The 101 data sets: selection criteria and baseline DASH performance 52
3.1 Introduction 53
3.2 Experimental 55
3.2.1 Data set treatments 55
3.2.2 Crystal structure solution 56
3.3 Results – the baseline DASH performance 62
3.4 Discussion 65
3.4.1 Dataset analysis 65
3.4.1.1 Space Group trends 65
3.4.1.2 DoF trends 66
3.4.1.3 Additional remarks 67
3.4.2 Baseline DASH performance 70
3.4.2.1 Simple crystal structures, DoF <14 72
3.4.2.2 Crystal structures of moderate complexity, 14 ≥ DoF ≤ 20 73
3.4.2.3 Complex crystal structures, 21 ≥ DoF ≤ 30 74
3.4.2.4 'Intractable' crystal structures, DoF > 30 76
3.4.2.5 Statistical analysis of the baseline DASH performance 77
3.5 Conclusions 83
3.6 References 85
4 Optimisation of DASH using irace 86
4.1 Introduction 87
4.1.1 The irace package 89

4.1.1.1 The irace input 90
4.1.1.2 The iteration 90
4.2 Experimental 92
4.2.1 The irace calculations 92
4.2.1.1 Configuring irace 92
4.2.1.2 irace runs 92
4.2.2 Evaluation 93
4.3 Results 94
4.3.1 irace results 94
4.3.2 Configuration evaluation 96
4.4 Discussion 100
4.4.1 Runs 1-10 102
4.4.2 Runs 11-14 103
4.4.3 Additional calculations 104
4.4.3.1 Compound A32 - Ibuprofen 104
4.4.3.2 Evaluation of the additional runs 105
4.4.3.3 The γ-carbamazepine exception 107
4.5 Conclusions 108
4.6 References 110
5 Optimised DASH – implementation of aggressive parameter settings 111
5.1 Introduction 112
5.2 Experimental 112
5.3 Results 112
5.4 Discussion 115
5.4.1 Success rate improvements 115
5.4.2 Gains in calculation times 118
5.4.3 Crystal structure solution accuracy 119
5.4.4 Statistical Analysis 120
5.5 Conclusions 125

5.6 References 125

6 Exploiting prior conformational knowledge 126

6.1 Introduction 127

6.1.1 Mogul and Mogul distribution bias (MDB) 128

6.1.2 Likely conformers as constrained starting models 131

6.2 Experimental 134

6.2.1 Mogul and MDB 134

6.2.2 Likely conformers as constrained starting models 134

6.3 Results 135

6.3.1 Mogul and MDB 135

6.3.2 Likely conformers 139

6.4 Discussion 139

6.4.1 Mogul and MDB – 'in-process' biasing 140

6.4.2 Likely conformers as constrained starting models – '*a priori'* approach to introducing conformational information 143

6.4.3 Statistical analysis 148

6.5 Conclusions 152

6.6 References 153

7 General conclusions 154

7.1 References 157

List of figures

Chapter 1

Figure 1.1. A flow chart of the SDPD process 5
Figure 1.2. A generic flowchart of GO methods implementation to SDPD..... 9
Figure 1.3. The molecular structures of verapamil (a) and tetracycline (b).. 14
Figure 1.4. A DASH specific workflow of Figure 1.2. 15
Figure 1.5. A pseudocode of the SA algorithm as implemented in DASH. 18
Figure 1.6. The number of molecular crystal structures deposited each year since 1990 in the CSD. 21
Figure 1.7. The mean and maximum of the total number of atoms in the asymmetric unit of molecular crystal structures deposited each year since 1990 in the CSD. 22
Figure 1.8.The mean and maximum total number of degrees of freedom in the asymmetric unit of molecular crystal structures deposited each year since 1990 in the CSD. 22
Figure 1.9. The molecular structures of the compounds listed in Table 1.2 24

Chapter 2

Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1 41

Chapter 3

Figure 3.1 A graph of the success rate (SR) vs DoF, based on the results published by Florence *et al.* (2005). 53
Figure 3.2 The crystal structure overlay of the CSD deposited crystal structure (in green) and a) the best DASH solution; and b) the DASH solution with χ^2 = 15.21. 62
Figure 3.3 A comparison of the relative space group distribution between the FDS and the CSD 66
Figure 3.4 The distribution of crystal structures plotted as a function of their DoF. 68
Figure 3.5 The powder X-ray diffraction data of B3..... 69
Figure 3.6 Overlay of the best DASH solution of B3 and the CSD deposited crystal structure. 69
Figure 3.7 Graphical representation of the baseline DASH success rate as a function of the total degrees of freedom (SA parameters used: T_0=0; CR = 0.02; N_1=20; and N_2=25)..... 71

Figure 3.8 The crystal structure conformation of A7: a) the best DASH solution, and b) the next best solution. 73
Figure 3.9 The crystal structure overlay of the best DASH solution for B41and its reference. 74
Figure 3.10 An overlay of the first and second DASH solutions of A34 (verapamil HCl). ... 76
Figure 3.11 The crystal structure overlay of the best DASH solution for B60 and its reference 77
Figure 3.12 The ELO model based on the total DoF.. 79
Figure 3.13 The ELO model based on the positional and torsional DoF. 80
Figure 3.14 An overlay of the observed and predicted SRs. 82

Chapter 4

Figure 4.1 The SA parameters tuning workflow 90

Chapter 5

Figure 5.1 Overlay of the best aggressive DASH solution and the reference crystal structure 117
Figure 5.2 The best baseline DASH solution of B56. 117
Figure 5.3 Crystal stucture overlay of the reference A40 crystal ctructure and: a) the best baseline DASH solution; b) the best aggressive DASH solution. 120
Figure 5.4 The aggressive ELO model based on the total DoF (solid green line). 121
Figure 5.5 Comparison of the default and aggressive ELO models based on the total DoF. 122
Figure 5.6 The aggressive ELO model based on the positional and torsional DoF. 123
Figure 5.7 Comparison of the default and aggressive ELO models based on the positional and torsional DoF.. 124

Chapter 6

Figure 6.1 The Mogul-derived distribution of the C6-C5-O2-C20 torsion angle based on the CSD entries. 129
Figure 6.2. The modal ranges of constraint/s as applied by Mogul. 130
Figure 6.3 The 'parameter bounds' window of DASH. 131

Figure 6.4 Conformer generator (0.9.3) workflow (CCDC, 2015) 132
Figure 6.5 The use of likely conformers as constrained starting models to solve crystal structures from powder diffraction data with DASH. 133
Figure 6.6 Hydrogen bond network of Ritonavir (form II). 141
Figure 6.7 An overlay of the A34 reference crystal structure and a DASH solution of χ^2 of 57. 143
Figure 6.8 Overlays of the reference crystal structures and the best conformer of: a) B56; b) A25; c) B53 and d) B61. 145
Figure 6.9 ELO models based on the results form a) $Mogul_{Default}$; b) $MDB_{Default}$; c) $Mogul_{Aggressive}$; and d) $MDB_{Aggressive}$. 149
Figure 6.10 An overlay of the four Mogul/MDB ELO models based on the results from the 51 tested compounds 150
Figure 6.11An overlay of all ELO models based on the total DoF (calculated against the results of the FDS). 151

List of Tables

Chapter 1

Table 1.1 Additional SDPD methods with an example of their application. 13
Table 1.2. A summary of SDPD crystal structures, selected on the basis of pharmaceutical and historical interest. 23

Chapter 2

Table 2.1 Compound names and corresponding CSD reference codes of the 101 previously-solved crystal structures, together with the code names used throughout this thesis. 37
Table 2.2 Summary of the PXRD data collection parameters used for Ritonavir and Lisinopril dihydrate.. 39
Table 2.3 Summary of used crystallographic software 47
Table 2.4 Hardware summary 47

Chapter 3

Table 3.1 Crystallographic information of the FDS as previously reported**.** 58
Table 3.2 A summary of the Pawley refinement details and baseline DASH performance against the FDS based on the 50 and 100 SA runs (1×10^7 moves). 63
Table 3.3 A summary of the Pawley refinement details and baseline DASH performance against the FDS based on the 500 SA runs (5×10^7 moves). 65
Table 3.4 Distribution of space groups within the FDS and the CSD. 66
Table 3.5 The Average SR of each DoF between 6 and 20. 74
Table 3.6 The Average SR of each DoF between 20 and 30. 75
Table 3.7 DoF composition of the intractable compounds. 77
Table 3.8 Regression analysis of the ELO *vs.* positional, orientational and torsional DoF. ... 79
Table 3.9 SR information of the crystal structures with 18 DoF. 81

Chapter 4

Table 4.1 Definitions the used irace related terms. 89
Table 4.2 The steps performed during a single irace iteration. 91
Table 4.3. The irace configuration 92
Table 4.4 An outline of the irace runs performed. 93
Table 4.5 A summary of the 14 compounds comprising the Evaluation set. 94

Table 4.6 Summary of the elite SA parameters configurations, as a result of the irace calculations. 95
Table 4.7 Summary of the SA parameters elite configurations, as a result of the irace calculations.. 96
Table 4.8 Evaluation of the highest ranked elite configurations of the irace runs 1-10. 97
Table 4.9 The evaluation of the elite configurations of the irace runs 11-14. 98
Table 4.10 The evaluation of the additional, semi-arbitrary test configurations 99
Table 4.11 A summary of the work carried out in Chapter 4. 102
Table 4.12 Assessment of a test configuration devised from the combinations of aggressive and default SA parameter values. 103
Table 4.13 The evaluation results of the first elite configurations of irace runs 11-14. 104
Table 4.14 The evaluation results of all elite configurations of irace runs 11-14. 104
Table 4.15 The evaluation of additional test configurations. 105
Table 4.16 The evaluation results of the additional test configurations. 106
Table 4.17 Summary of the best performing SA parameter configurations and their average SR. 107
Table 4.18 Comparison between the results the current DASH algorithm and the adjusted DASH algorithm for the purpose of A38. 108

Chapter 5

Table 5.1 A summary of the SRs achieved with the six aggressive SA parameter configurations, against the FDS, based on the 50 and 100 SA runs. 113
Table 5.2 A summary of the SRs achieved with the six aggressive SA parameter configurations, against the FDS, based on the 500 SA runs 115
Table 5.3 The average SR of the structural complexity groups. 116
Table 5.4 A summary of the performance improvements of the 0.27; 73; 56 and 0.27; 73; 61 configuration 117
Table 5.5 Default and aggressive SRs achieved for compounds B44-B47. 118
Table 5.6 Calculation times for B45-B47 as a function of the χ^2 multiplier = 5. 119
Table 5.7 Regression analysis of the aggressive ELO *vs*. positional, orientational and torsional DoF. 122
Table 5.8 Calculated and experimental SRs of the compounds with 28 DoF. 124

Chapter 6

Table 6.1 FDASH runs performed with the selected 10 compounds. 135
Table 6.2 Mogul and MDB results based on the 50 and 100 SA runs using default and aggressive parameters. 136
Table 6.3 Mogul and MDB results based on the 500 SA runs using default and aggressive parameters 138
Table 6.4 Results of the conformer ensembles evaluation for compounds requiring up to 100 SA runs 139
Table 6.5 Results of the conformer ensembles evaluation for compounds requiring 500 SA runs (with the use of default SA parameters) 139
Table 6.6 Evaluation of the accuracy of the conformer ensembles against the reference crystal structures of the 10 selected compounds. 144
Table 6.7 Results of the preliminary DASH runs 147
Table 6.8 Regression analysis of the default Mogul and MDB ELO models *vs.* positional, orientational and torsional DoF. 148
Table 6.9 Regression analysis of the aggressive Mogul and MDB ELO models *vs.* positional, orientational and torsional DoF. 148

List of equations

Equation 1.1 6
Equation 1.2 16
Equation 1.3 18
Equation 1.4 19
Equation 3.1 78
Equation 3.2 78
Equation 3.3 78
Equation 3.4 80
Equation 4.1 90
Equation 4.2 90
Equation 4.3 92
Equation 4.4 101
Equation 5.1 120
Equation 5.2 122

List of Appendices

Appendix A.

Powder X-Ray diffraction data and DASH files of the 101 datasets, necessary to conduct the calculations performed.

Appendix B.

Graphs of χ^2 *vs.* SA moves of the 10 compounds used to evaluate the introduction of prior conformational knowledge in the form of constrained stating models.

The appendices are supplied in a digital form on the enclosed CD and memory card.

Frequently used abbreviations

AGG	Aggressive simulated annealing parameter settings
CPU	Central Processing Unit
CR	Cool Rate
CSD	Cambridge Structural Database
CSD code	Cambridge Structural Database crystal structure reference code
DEF	Default simulated annealing parameter setting
DoF	Degree(s) of Freedom
ELO	Empirical Log of the Odds
FDS	Full Data Set
FF	Fully Flexible
FOM	Figure of Merit
GA	Genetic Algorithm
GO	Global Optimisation
PXRD	Powder X-Ray Diffraction
RMSD	Root-Mean-Square Deviation
SA	Simulated Annealing
SDPD	Structure Determination from Powder Diffraction Data
SR	Success Rate

1 Introduction

Knowledge of the three-dimensional structure of crystalline materials is crucial to understanding their chemical, biological, and physical properties. Whilst numerous analytical techniques, including thermal and spectroscopic methods, can give valuable information about crystalline and amorphous materials, currently only X-ray diffraction techniques are capable of routinely, directly determining the three-dimensional structure of crystalline materials. Therefore, X-ray techniques are key tools used in areas such as chemistry, pharmacy, biology, material science, mineralogy, and physics. For example, in the pharmaceutical industry, knowledge and control of the drug and excipient properties are essential in bringing a drug to market.

The preferred X-ray diffraction method for full crystal structure determination is single-crystal X-ray diffraction, which is considered the 'gold standard' if a suitably large (*ca.* 60 μm in all dimensions), high quality single-crystal (SX) of the molecule under study can be obtained. In the event that such a crystal cannot be easily grown, powder X-ray diffraction (PXRD) is, increasingly, a suitable alternative method. Of particular interest are active pharmaceutical ingredients (APIs) which are normally processed as polycrystalline powders and which can not only exist in multiple polymorphic forms but also as hydrates, solvates, salts and co-crystals. It is perhaps not surprising, therefore, that PXRD has been found to be the most frequently used analytical tool when studying pharmaceutical materials (Chieng *et al.*, 2011).

The versatility of PXRD is demonstrated by its use throughout the different stages of the drug manufacturing lifecycle (Brittain, 2001; Ivanisevic *et al.*, 2010; Randall *et al.*, 2010). Its importance as a 'fingerprint' with which to identify specific crystallographic phases is evidenced by its central role in high-throughput physical form screening and its use (either in diagrammatic form, or as a series of reflection positions) in patents designed to protect physical forms. These aspects have been reviewed comprehensively elsewhere (Florence, 2009; Lemmerer *et al.*, 2011; Morissette *et al.*, 2004). Nowadays, crystal structure determination from powder diffraction data (SDPD) can be considered as a routine, though not always straightforward, approach that has the advantage of dealing directly with a bulk polycrystalline sample, which may consist of multiple crystallographic phases. This is in contrast to single-crystal diffraction, where a single crystal grown from a solution is not necessarily representative of the bulk material being handled in the pharmaceutical workflow. It is worth noting that the overall number of structures (not confined to pharmaceutical materials) solved using PXRD accounts for only a very small percentage of the total number of solved crystal structures; this is a direct consequence of the ease with which structures can be solved when

single crystals are available and when such crystals are available, they are always the first port of call. For those structures where single crystals cannot be obtained, SDPD has become a key tool for fully populating the crystallographic structural landscape.

1.1 Solving crystal structures from powder diffraction data

To fully characterise a crystal structure model, the direction, magnitude and phase angle for each reflection need to be established. Whilst the direction and amplitude can be relatively easily identified in a diffraction experiment - the former by establishing the Miller indices of diffracted beams and the latter based on their intensity - the relative phase is lost. In crystallography this is referred to as 'the phase problem' and for the vast majority of small molecule crystal structures it can be considered to be a solved problem, given accurate structure factors to atomic resolution. The foundations of the computational direct phasing methods approach to crystal structure determination, laid down in the 1950s and developed over the following decades, has led to computer programs which (when dealing with good quality SX lab-based diffraction data) can easily solve structures containing in excess of 150 atoms in the asymmetric unit. A significant factor in this success is the fact that modern diffractometers can collect SX data sets to near 100% completion to high (*ca*. 0.8 Å) resolution, providing typically thousands of reflections upon which the phasing process can operate.

1.1.1 The intrinsic problems of PXRD

The collapse of the three dimensions of single-crystal diffraction into the one dimension of a powder diffraction pattern is undoubtedly the major issue facing SDPD. As a consequence, accidental reflection overlap *(i.e*. the overlap of diffraction contributions from non-symmetry related reflections having very similar *d*-spacings) is observed, especially with increasing diffraction angle, and the number of reflections that can be observed as individual peaks in the PXRD pattern is significantly reduced. This issue leads to difficulties with the assignment of individual reflection intensities, a problem further compounded by the fact that well-determined reflections frequently do not extend close to atomic resolution, due to a combination of factors such as the scattering (form) factor fall off, the temperature factor and the Lorentz-polarisation factor. Thus in a PXRD pattern obtained in the lab from a typical small molecule structure only a few hundred reflections will be observed (*c.f.* SX, a few thousand) and as such, conventional direct methods of structure determination do not perform well on such limited data. Strategies such as the use of variable count time (VCT) data collection and

low-temperature data collection can help (Madsen and Hill, 1994; Shankland *et al.*, 1997), but are seldom sufficient to permit the application of *conventional* direct methods.

Unique to the SDPD is also the occurrence of preferred orientation, where the packing of the crystallites in the polycrystalline sample is non-random, thus affecting the observed intensities in the diffraction pattern. Different approaches exist to address this issue, primarily the use of data collection in transmission capillary mode and the use of correction terms, such as the March-Dollase (Dollase, 1986) and spherical harmonics (Sitepu *et al.*, 2005) corrections which are employed, if needed, during the structure solution and refinement stages (see 1.1.2 for the SDPD workflow).

Overall, the SDPD process remains significantly more challenging than its single-crystal counterpart.

1.1.2 The SDPD workflow

The process of SDPD consists of a number of sequential steps, the success of each enabling the execution of each subsequent step and therefore the success of the entire crystal structure solution (Figure 1.1). This dependency exists from as early as sample collection: a sample containing impurities or that is poorly crystalline may prove impossible to solve.

Given a well prepared sample, the work flow follows the steps given in Figure 1.1. The experienced crystallographer will make a decision about the appropriate method (shown in the Figure's middle column) to be selected at each stage; these will be dependent on the complexity of the sample and the quality of the final result required.

Many comprehensive reviews of SDPD methodologies have now been published (Cerny and Favre-Nicolin, 2007; Datta and Grant, 2004; David and Shankland, 2008; Harris, 2012; Tremayne, 2004) and in particular, the IUCr Monograph on Crystallography "Structure Determination from Powder Diffraction Data" (Shankland *et al.*, 2002) provides a 'powder sample to refined crystal structure' view of the process. Hence, only the actual structure solution step and its associated methods will be further discussed here.

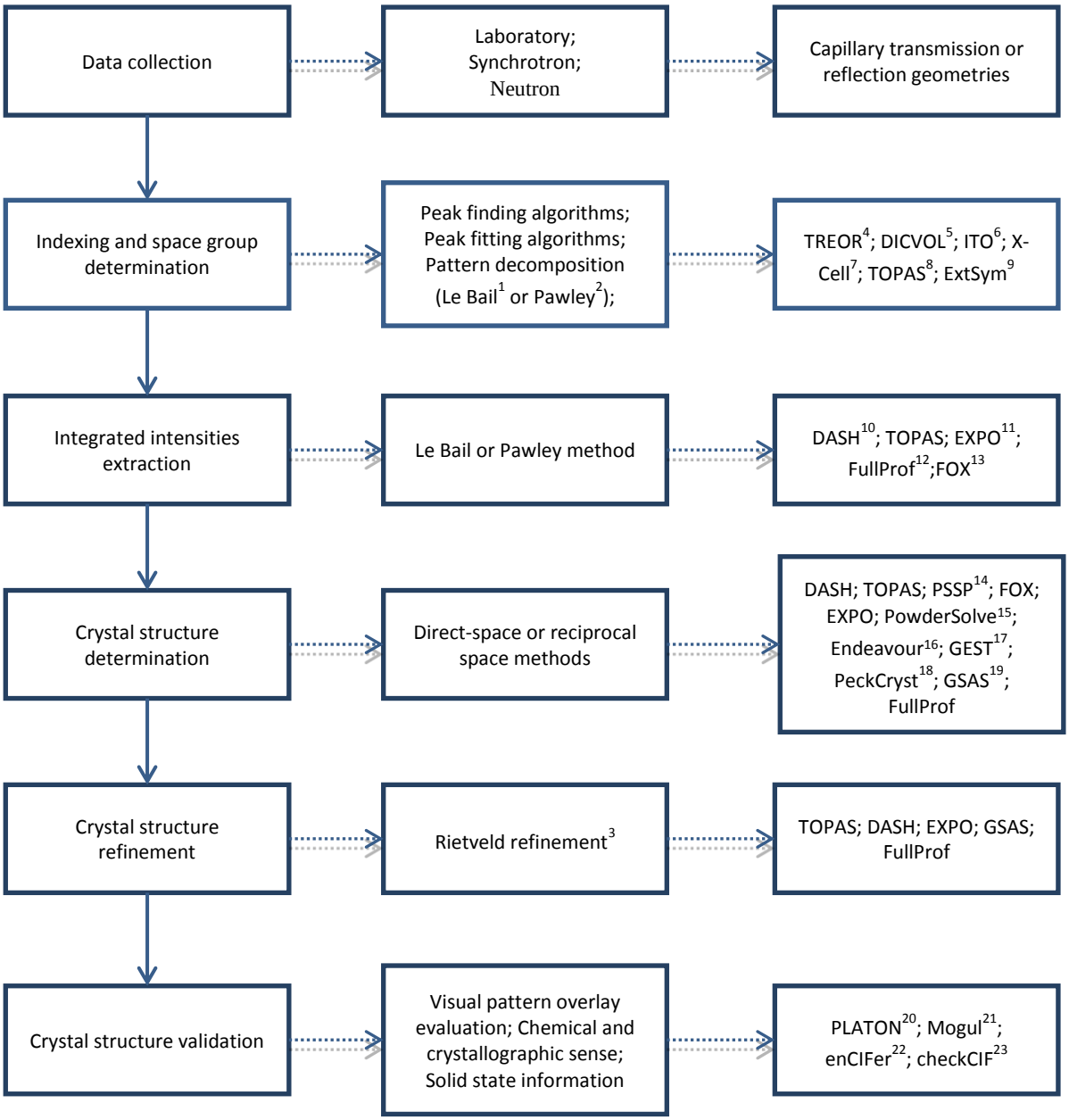


Figure 1.1. A flow chart of the SDPD process (left column); examples of the applied methodologies (middle column); and software where those have been applied (right column). 1 (Le Bail *et al.*, 1988); 2 (Pawley, 1981); 3 (Rietveld, 1969); 4 (Werner *et al.*, 1985); 5 (Boultif and Louer, 1991); 6 (Visser, 1969); 7 (Neumann, 2003); 8 (Coelho, 2003); 9 (Markvardsen *et al.*, 2008); 10 (David *et al.*, 2006); 11 (Altomare *et al.*, 2011); 12 (Rodriguez-Carvajal, 1993); 13 (Favre-Nicolin and Cerny, 2002); 14 (Pagola and Stephens, 2010); 15 (Engel *et al.*, 1999); 16 (Putz *et al.*, 1999); 17 (Feng and Dong, 2007); 18 (Feng *et al.*, 2009); 19 (Larson and Von Dreele, 1994); 20 (Spek, 2003); 21 (Bruno *et al.*, 2004); 22 (Allen *et al.*, 2004); 23 (IUCr, 2014)

1.2 Methods for SDPD

Early attempts to solve crystal structures from powder diffraction data were based on the conventional single-crystal approaches (such as Direct, Patterson and maximum entropy methods), adapted to address the information loss observed with powder X-ray diffraction (Debets, 1968; Zachariasen and Ellinger, 1963).

Additionally, trial-and-error approaches to crystal structure determination were developed. These approaches, however, require knowledge of the connectivity of the molecule in order to construct an initial model. A logical development of these trial-and-error methods, greatly enabled by advances in computing power, was the stochastic methods (typified by simulated annealing and genetic algorithms) that now predominate.

A brief description of some of the widely used computational techniques is given below, whilst examples of their applications can be found in Table 1.2 of Section 1.5.

1.2.1 Single-crystal structure solution methods adapted for powder diffraction data

1.2.1.1 Conventional and modified direct methods

Conventional direct methods of structure solution derive the crystal structure from an electron density map (assuming the use of X-rays), which is a Fourier transform of a set of phased structure factors (Equation 1.1).

$$\rho(xyz) = V^{-1} \sum_{h} \sum_{k} \sum_{l} |F| cos\{2\pi(hx + ky + lz) - \phi_{hkl}\} \qquad \text{Equation 1.1}$$

where $\rho(xyz)$ is the electron density; V is the volume of the unit cell; $|F|$ are the structure factors magnitude, h,k,l are the Miller indices and ϕ_{hkl} is the phase angle.

As a result, direct methods do not require prior knowledge of the connectivity of the structure; the structure is derived from the electron density. However, the success of the phasing process is greatly dependent upon the accuracy and resolution of the structure factors and, in general, it requires accurate structure factors collected to approximately atomic resolution (*ca*. 0.9Å). For typical molecular organic crystal structures with relatively large, low symmetry unit cells, accidental reflection overlap and the form-factor fall-off conspire to make this difficult to achieve. As such, the use of conventional direct methods is best suited for higher symmetry, strongly scattering samples such as inorganic and organometallic compounds.

In order to achieve success with more complex powder diffraction problems, direct methods have been considerably modified, primarily by recycling any structural fragments obtained from phasing into the intensity extraction stage, in order to improve the accuracy of the extracted structure factors. Other algorithmic developments (*e.g*. related to the chemical interpretation of low-resolution density maps) have added to this, resulting in SDPD computer

programs that are capable of solving complex molecular organic crystal structures; see, for example, EXPO2011 (Altomare *et al.*, 2011) which takes advantage of a resolution bias correction algorithm (RBM) and a weighted least-squares procedure (wLSQ). In the later version (EXPO2012), the COVMAP procedure (Altomare *et al.*, 2012) was added to modify and improve the models provided by the direct method procedure.

1.2.1.2 Patterson methods

The Patterson function (Patterson, 1934), whilst a Fourier series, is calculated using the squared structure factors $|F_{hkl}|^2$, thus obviating the requirement for phase information. Patterson maps containing *interatomic vectors* rather than atomic position maps are used to achieve the crystal structure determination. These vectors are drawn between all atoms in the crystal structure, and their heights are proportional to the sum of the atomic numbers of the contributing atoms. As such, the highest peaks are those between the heaviest atoms in the structure, allowing the heavy atoms to be located precisely. It is therefore no surprise that Patterson methods are particularly well suited to dealing with structures containing heavy atoms. The main disadvantage is that the map contains N^2 peaks, making interpretation difficult for structures of any reasonable complexity. Recently, a Patterson-function tangent formula was successfully used in combination with direct methods to solve the crystal structures of a number of organic compounds (Rius, 2011; Rius *et al.*, 2011).

1.2.1.3 Charge flipping

Charge flipping (CF) is a relatively new (Oszlanyi and Suto, 2004) but very promising method of structure solution. It has been widely utilised in single-crystal structure determination and implemented in a number of single-crystal software suites. Unlike direct methods, the algorithm does not depend directly on atomicity (Oszlanyi and Suto, 2008) and further differs from traditional direct methods as it is not reliant on probabilistic phase relations.

The process starts with the assignment of random phases to a set of experimental structure factor amplitudes, which is then subject to Fourier transform, generating an electron density map. Then, a positive density threshold is assigned and all points with density below this threshold are given a phase with an opposite sign. Fourier transform of this modified density yields a new set of structure factors whose phases are then combined with experimental amplitudes to generate new density map. This iterative cycle continues until such point as a likely crystal structure can be identified from the phased map.

The main advantage of this method is that no prior knowledge of the space group is required. The crystal structure solution is normally carried out in *P*1 symmetry, and the correct space group is derived at the end of the crystal structure solution. However, the CF algorithm still requires near atomic resolution, which once again is the limiting factor for its full utilisation with powder X-ray diffraction data.

The charge flipping algorithms have been implemented in a range of software packages such as Superflip (Baerlocher *et al.*, 2007), TOPAS (Coelho, 2003) and Jana2006 (Petricek *et al.*, 2014) and has been extensively utilised in the work of McCusker on zeolites (Pinar *et al.*, 2011; Xie *et al.*, 2011a; Xie *et al.*, 2011b) More recently Jung *et al* (2014) have shown that organic compounds containing only light-atoms pose a challenge to the charge flipping algorithm (as implemented in Superflip), which performs well with inorganic compounds.

1.2.1.4 Maximum entropy

The combined maximum entropy/log-likelihood gain approach is another alternative to direct methods. It was first proposed by Bricogne (1984) for the structure solution from single crystal data, and later extended to powder diffraction data (Bricogne, 1991). The powder-specific method divides peaks into overlapping and non-overlapping sets, each of the group's intensity are calculated and normalised to give initial structure factors. The origin is defined by fixing the phases of a number of non-overlapping origin-defining reflections. These reflections establish the initial basis set, which is then expanded in the structure solution process to form a "phasing tree" of possible basis sets. The phases of the reflections in each basis set are used as constraints in a subsequent entropy maximisation procedure. A likelihood function is used to evaluate the most probable basis set, which is then further expanded by the addition of yet more reflections. The process is terminated once sufficient structure factors have been phased to allow a good quality electron density map to be generated.

The only program using this combined maximum entropy/log-likelihood gain algorithm is MICE (Gilmore and Bricogne, 1997), although other programs utilise maximum entropy in a different role *e.g.* RIETAN-2000 (Izumi, 2004). The RIETAN-2000 program was recently used to solve the structures of $Ca_{1-x}Bi_xMn_{1-y}V_yO_{3-\delta}$ solid solutions where $\delta \leq x = y \leq 0.08$, from powder X-ray data (Huang *et al.*, 2008). Another recent example is the crystal structure solution of the medium-sized pharmaceutical prednisolone succinate (Nishibori *et al.*, 2008), which was achieved by combination of genetic algorithm and a maximum entropy approach.

1.2.2 Direct-space methods

In the direct space approach, the position, orientation and conformation of a 3D structural model of the molecule to be determined is adjusted within the unit cell as derived from the PXRD data. The problem of finding the best agreement between the observed and calculated structure factors is that of locating the global minimum on an agreement factor (*e.g.* R_{wp} or χ^2) hypersurface, whose dimensionality equals the number of structural variables in the problem. Such hypersurfaces encompass a multitude of stationary points and locating the global minimum is a non-trivial exercise which requires the application of global optimisation (GO) algorithms. These methods require knowledge of the molecular connectivity, in order to create a 3D model of the molecule under study.

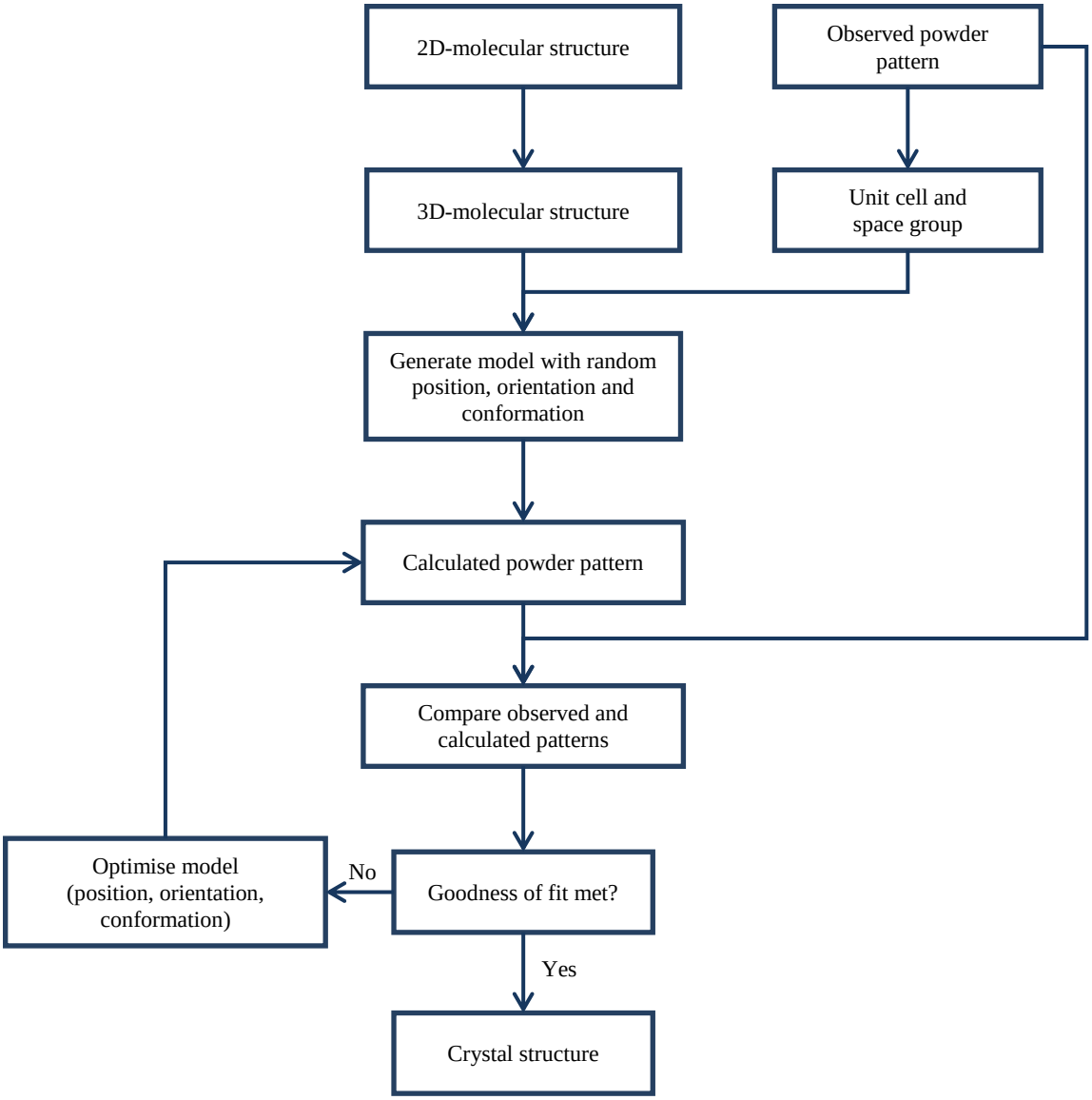


Figure 1.2. A generic flowchart of GO methods implementation to SDPD.

Molecular models are easily constructed with software such as MarvinSketch (ChemAxon, 2011), or ChemDraw (Cambridgesoft, 1985-2015). Cambridge Structural Database (CSD) deposited structures, closely related to the molecule under study, can also be an excellent source for a reliable starting models.

Regardless of the way the starting model is generated, it is crucial to ensure its underlying molecular geometry is chemically sensible and follows well defined trends. This will maximise the chances of the GO algorithm reaching a successful solution, and significantly reduce the necessary work at the crystal structure refinement stage. The CSD, and particularly Mogul, can be employed to perform such validation. Quantum mechanical packages, such as MOPAC (Stewart, 1990), Firefly (Granovsky), QuantumEspresso (Paolo *et al.*, 2009), Gaussian (Frisch *et al.*, 2009) etc., can also be employed. It has to be mentioned that the calculation time with such packages increases very quickly with increased number of atoms and required accuracy of the model, but can be considered to be time well spent in order to optimise the chances of successful crystal structure solution.

It should be apparent from the above flowchart that the key element is the model optimisation. There are a great many algorithms for performing this optimisation, ranging from simplistic / exhaustive (*e.g.* grid search) through to complex / stochastic / deterministic. The focus of this following section is on the two main approaches that have been most widely used in SDPD: simulated annealing (SA) and genetic algorithms (GA). Simulated annealing in the context of the DASH program is discussed in detail in Section 1.4.

1.2.2.1 Simulated annealing

The SA was first developed in the early 1980s (Kirkpatrick *et al.*, 1983) (Cerny, 1985; Smith *et al.*, 1983) and since has been extensively employed in optimising manufacturing design (Renzi *et al.*, 2014; Sun and Huang, 2012; Yusup *et al.*, 2012), the traveling salesman problem (Ye and Rui, 2013), multi-dimensional assignment problems (Clemons *et al.*, 2004) and many other areas (Li and Liu, 2013; Mohagheghian *et al.*, 2015; Moschakis and Karatza, 2015; Radu and Vintan, 2013; Rahimian *et al.*, 2015).

In the context of SDPD, SA was first introduced by Deem and Newsam (Treacy *et al.*, 1989) for zeolites, and is currently the most widely used optimisation method (Shankland and David, 2002). The process is efficient, easy to use, and it has relatively few variables, which can be set automatically.

The approach is an implementation of a Monte Carlo method, which is inspired by analogy to the physical annealing process[1]. In the structure solution stage, the crystallographic model is initially "melted", allowing all possible positions, orientations and conformations of the molecule in the unit cell to be adopted. Slow lowering of temperature (T) follows, during which the system assumes its true crystal structure (equivalent to the global minimum).

For the SA to work properly, an appropriately high starting temperature needs to be used to simulate the melting phase. However, unnecessary high values of T would require additional SA moves for the global minimum to be reached. Equally important is the rate at which the T is reduced, the "cooling rate", as fast T reduction is likely to trap the algorithm in a local minimum (the equivalent of quenching), but a slow cooling may unnecessarily prolong the process of locating the global minimum.

Different approaches to SA have been developed. By way of example the SA developed by Andreev and Bruce (1998) and that implemented in ENDEAVOUR (Putz *et al.*, 1999), reduce the T at a pre-set rate, whilst in DASH (David *et al.*, 1998), fluctuation of the cost function regulates the reduction of temperature, thus allowing a more in-depth exploration of regions of higher function value.

Other programs based on the SA algorithm include, TOPAS (Coelho, 2003), FOX (Favre-Nicolin and Cerny, 2002) and PowderSolve (Engel *et al.*, 1999). Simulated annealing has now been also integrated in EXPO2011 (Altomare *et al.*, 2011), which typically used only *ab initio* methods.

Due to its ease of use and availability, many crystal structures have been solved with the use of SA (see Table 1.2). One of the example from the table worth noting is the chlorothiazide dimethylformamide (1/2) solvate (Fernandes *et al.*, 2007), with 6 fragments in the asymmetric unit cell and a total of 42 optimisable parameters during the SA calculations, is still one of the most complex compounds of pharmaceutical interest solved from powder diffraction data to date.

1.2.2.2 Genetic algorithm

As with all evolutionary algorithms, the genetic algorithm (GA) is an optimisation algorithm with a biological basis. It is based on Darwinian Theory and inspired by naturally accruing

[1] e.*g.* Melting of a metal and its subsequent slow cooling to an ordered state.

processes such as inheritance, mutation and natural selection. A recent review by Paszkowicz (2009) gives a detailed listing (with further references) of the GA application in material sciences, chemistry and physics of materials, including applications in crystallography.

In the context of SDPD, GA is also widely used. Here the optimisable parameters are considered as chromosomes, the search for the structure solution is equivalent to a search for an individual with best fitness, and the cost function is related to the fitness of each individual.

The algorithm starts with an initial population of molecules. Natural selection is employed to select the individuals with the best fitness and the genetic operations of crossover and mutation are performed, with the population evolving towards the point of global minimum. The algorithm is terminated if a pre-set value of the goodness-of-fit or the maximum number of generations is reached.

Early applications of GA include the crystal structure determination of ortho-thymotic acid (Kariuki *et al.*, 1997), ibuprofen (Shankland *et al.*, 1998) and fluticasone propionate (Kariuki *et al.*, 1999). More recently the work of Harris and co-workers applies the genetic algorithms to a range of organic molecules, including l-arginine (one of the few natural amino acids with previously undetermined crystal structure) (Courvoisier *et al.*, 2012); another interesting example is the structure of hexaketocyclohexane octahydrate, characteristic with its unusually high density due to a large number of hydrogen bonding (16 hydrogen-bond donors and 14 hydrogen-bond acceptors in the unit cell) (Lim *et al.*, 2011).

1.2.3 Other methods of SDPD

A number of other computational methods, which have been applied to SDPD, but did not merit a detailed discussion, have been summarised in Table 1.1.

Table 1.1 Additional SDPD methods with an example of their application. HEWL = hen egg-white lysozyme

Method	Examples	Reference
Single-crystal type methods		
Direct method sum function	XLENS	(Rius, 1999)
Anomalous scattering	HEWL:cisplatin co-crystal	(Tanley *et al.*, 2012)
Isomorphous replacement	HEWL	(Basso *et al.*, 2010)
Molecular replacement	Clarithromicyn I	(Noguchi *et al.*, 2012)
Direct-space methods		
Grid Search	P-RISCON	(Masciocchi *et al.*, 1994)
Parallel tempering	FOX	(Favre-Nicolin and Cerny, 2002)
Particle swarm	PeckCryst	(Feng *et al.*, 2009)
Local minimisation	Famotidine	(Shankland and David, 2002)
Simplex (Semi-global)	DASH	(David *et al.*, 2006)
Hybrid methods		
Hybrid Monte Carlo	Capsaicin	(Johnston *et al.*, 2002)
Hybrid big bang big crunch	EXPO2013	(Altomare *et al.*, 2013b)

1.2.4 Summary of the SDPD methods

In a very brief summary, the methods for crystal structure solution can be very broadly divided into two subsets: 1) reciprocal space methods, which do not require prior information of the structure under study, but are limited by the near Ångström-resolution data requirement; and 2) direct space methods, for which the knowledge of the molecular connectivity is a prerequisite; which however, are not faced with the phase problem of the reciprocal space.

Organic molecular crystals are an example of a challenging problem for the traditional direct method, which is well 'suited' to the nature of GO algorithms.

1.3 Powder diffraction and crystal structural complexity

The suitability or success of each of these methods described in the previous section is often judged in terms of the complexity of the molecules on which they can be brought to bear. As such it is important to discuss the different ways in which complexity is defined for SDPD. With most single-crystal-like methods the complexity of the structure under study is described by the number of non-hydrogen atoms in the asymmetric unit, as this is the number of atomic positions being determined during the crystal structure solution step. This description is certainly appropriate for DM-based methods of SDPD but is less suitable for GO-based methods, where the number of parameters being optimised is a more appropriate measure. The total number degrees of freedom (DoF) for a molecular crystal structure is the sum of the external DoF (3 positional and 3 orientational for each the independent fragment in the asymmetric unit) and the internal DoF (the number of flexible torsion angles for each

independent fragment in the asymmetric unit). Figure 1.3, which shows the molecular structures of tetracycline and verapamil, illustrates how these two complexity descriprtions can differ significantly depending upon the molecules under consideration.

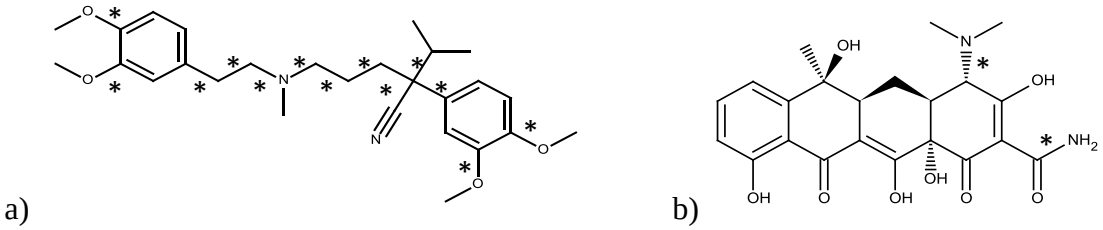


Figure 1.3. The molecular structures of verapamil (a) and tetracycline (b). Internal DoF (torsion angles that are free to rotate) are indicated by an asterisk. Assuming that the ring conformations are known, the structure of tetracycline (32 non-H atoms) can be described by 6 external DoF plus 2 internal DoF whilst that of verapamil (33 non-H atoms) requires 6 external DoF plus 13 internal DoF. Thus, tetracycline is the much simpler problem to tackle using global optimisation as fewer variables need to be determined.

1.4 SDPD and DASH

DASH is a computer program for solving crystal structures from powder X-ray diffraction data, which is optimised to deal with molecular materials. The first use of the methodology implemented in DASH dates from 1998, with the SDPD of three previously unknown crystal structures from synchrotron PXRD data: capsaicin, thiothixene and promazine hydrochloride (David *et al.*, 1998). All three structures were complex relative to other powder structures in the literature at that time. By way of example, approximately three quarters of the crystal structures deposited in the CSD at the time had fewer atoms in the asymmetric unit than thiothixene. DASH itself was released in 1999 and two key publications describe the program (David *et al.*, 2006) and its range of applicability (Florence *et al.*, 2005). The development of distributed computing versions of DASH has also been reported [MDASH, (Griffin *et al.*, 2009b); GDASH (Griffin *et al.*, 2009a)].

The workflow of DASH is similar to other programs employing GO algorithms. An overview is given in Figure 1.4 and the specific details relevant to this thesis are discussed below:

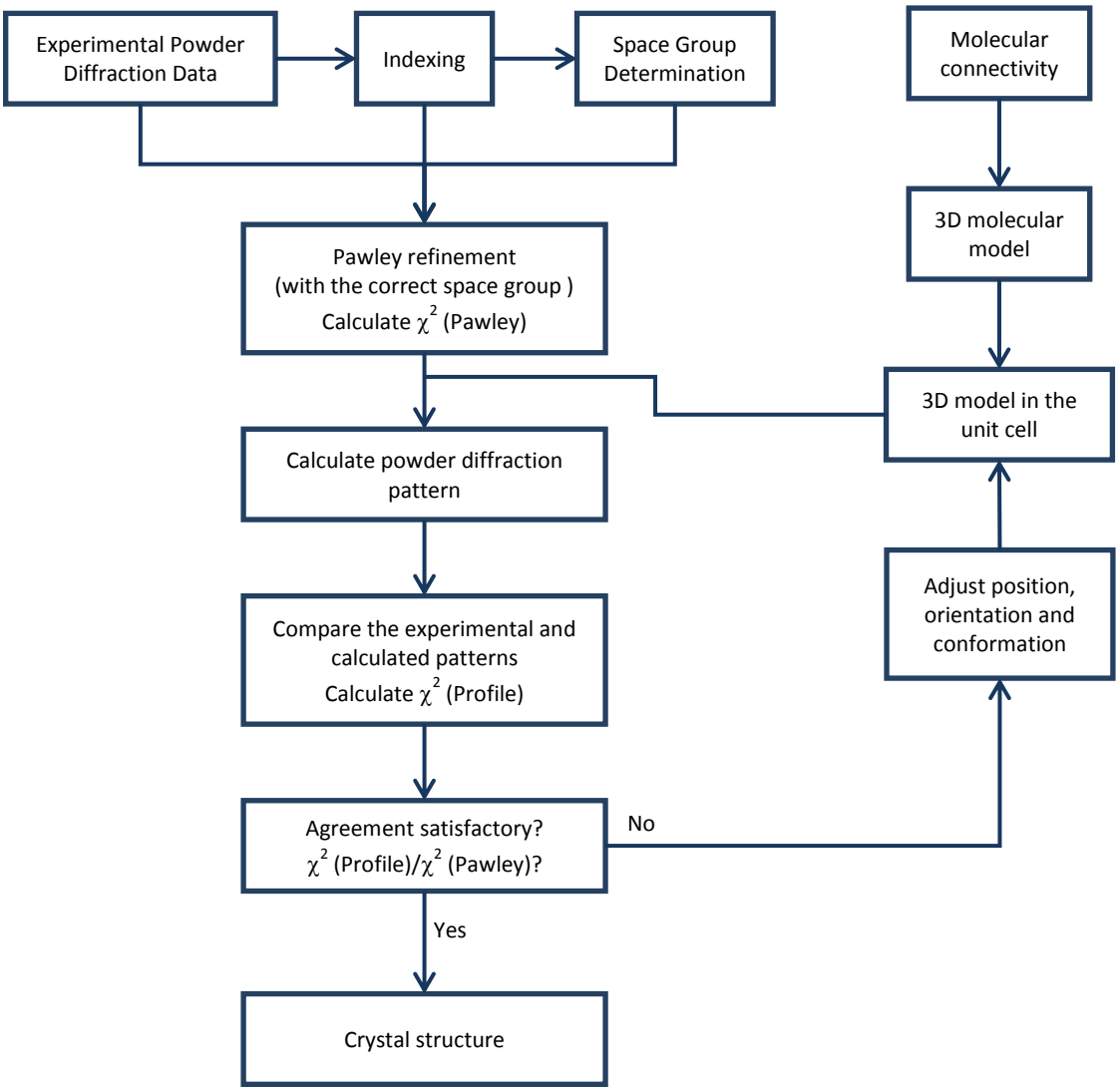


Figure 1.4. A DASH specific workflow of Figure 1.2.

By default DASH utilises DICVOL06 (or previous versions 91; 04; (Boultif and Louer, 1991, 2004) for the indexing of the powder X-ray diffraction pattern. However, DASH can also interface to other indexing programs and in fact improve their performance by providing accurate peak positions. During the Pawley refinement, the unit cell, peak shape and background parameters, together with the reflection intensities are optimised in a least-squares procedure. The resulting optimised values are used as prerequisites for the space group determination and crystal structure solution steps.

The Pawley fitting capability of DASH is used in two capacities. Firstly, a Pawley fit in a holosymmetric space group of the crystal structure under study (*e.g. P*2 for monoclinic cell, *P*222 for orthorhombic, *P*6 for hexagonal etc.) returns a set of extracted reflection intensities that are used by the ExtSym (Markvardsen *et al.*, 2008) program in order to determine the extinction symbol of the crystal structure under study. The results are presented as a ranked list

of log-probabilities of the extinction symbols belonging to the crystal system. The most probable space group corresponding to this extinction symbol is then selected by the user, taking into account factors such as observed space group frequencies and chirality.

A second Pawley refinement is then performed in the chosen space group in order to return a set of reflection intensities against which the calculated intensities derived from the trial crystal structures can be quickly compared in order to assess their correctness; this is discussed further below.

The goodness of fit of this Pawley refinement is given as a Pawley profile χ^2 (Equation 1.2) whose value serves two main purposes: a) as a measure of the best possible fit of the observed powder diffraction data when reflection intensities are allowed to refine as independent variables; and b) as a measure against which the profile χ^2 of the best structure returned by the simulated annealing stage can be compared. A crystal structure is considered to be worth examining carefully when it exhibits a favourable χ^2 ratio, typically in the range between 2 to 10, though the lower the better.

$$\chi^2 = \frac{\sum_i w_i (y_{i(obs)} - y_{i(calc)})^2}{N - P + C} \qquad \text{Equation 1.2}$$

where *N* is the number of observations, *P* is the number of parameters and *C* is the number of constraints.

As described above, global optimisation methods require the use of a 3D structural model whose parameters (position, orientation, and conformation) need to be determined in order to solve the crystal structure. For use in DASH, a 3D model of relative atomic coordinates is converted into the form of a Z-matrix, where the position of each atom is determined by the bond lengths, bond angles and torsion angles of the three preceding atoms. The 3D connectivity of the model is automatically detected and any flexible torsion angles are flagged as optimisable parameters (unless otherwise specified by the user). It is important that the input molecular model is as accurate as possible in terms of the core molecular geometry; the more accurate the model, the more likely it is that the crystal structure solution will be obtained. Generally, models obtained from (or derived from) the CSD or from high-level molecular modelling calculations are most suitable. These models can be compared to crystal structures in the CSD using Mogul, prior to their use within DASH.

The SA algorithm (as described in Section 1.2.2.1) implemented in DASH introduces a number of distinctive features: a novel cost function, an adaptive cooling rate and an efficient parameter space search. It is the combination of these features which facilitates the routine determination of complex structural problems.

The SA has five user definable parameters which control its performance[2]: the starting temperature (T_0), N_1 and N_2 (the values of which determine the number of steps made at each temperature), the cooling rate (CR) and the maximum number of SA moves allowed for each SA run. Currently, an automatic setting of T_0 ($T_0 = 0$) is used as the default. An algorithm determines a problem-specific value of T_0, such that virtually all uphill moves are allowed at the outset of the SA, *i.e.* the structure is "melted". This temperature is then reduced according to the cooling rate, which is discussed further below.

The role of N_1 and N_2 can be illustrated with the use of the pseudo code in Figure 1.5 for the case of a simple hypothetical molecule with one rotatable torsion angle (a total of DoF=7).

[2] For the purposes of this work, the DASH performance is defined by three terms - the success rate of crystal structure solution, the minimum number of SA steps required to achieve a correct crystal structure and the overall structure solution time.

1.	x	y	z	φ_1	φ_2	φ_3	τ

2. Set an initial T, to allow all uphill moves to be accepted

3. Sequentially

   Adjust x; evaluate χ^2
   Adjust y; evaluate χ^2
   Adjust z; evaluate χ^2
   Adjust φ_1; evaluate χ^2
   Adjust φ_2; evaluate χ^2
   Adjust φ_3; evaluate χ^2
   Adjust τ; evaluate χ^2

   } Repeat N_1 times

4. Adjust the step length for each parameter

5. Return to step 3; repeat N_2 times

6. Apply the cool rate to reduce T

Figure 1.5. A pseudocode of the SA algorithm as implemented in DASH. This specific example has three positional variables (x, y and z), three orientational (φ_1, φ_2 and φ_3) DoF and one torsion angle (τ).

The SA algorithm begins by assigning random values of the seven optimisable parameters, and by automatically calculating the optimal initial temperature. Then, the agreement between the calculated and observed intensities (from the Pawley refinement) is evaluated with the use of the integrated intensities χ^2 (Equation 1.3).

$$\chi^2 = \sum_h \sum_k [(I_h - c|F_h|^2)(V^{-1})_{hk}(I_k - c|F_k|^2)] \qquad \text{Equation 1.3}$$

where I_h and I_k are the extracted intensities from the Pawley refinement of the diffraction pattern, which have been corrected by the use of Lorentz-polarisation; V_{hk} is the covariance matrix from the Pawley refinement; c is a scale factor; and $|F_k|$ and $|F_h|$ are the structure factor magnitudes calculated from the trial structure.

This cost function speeds up the figure of merit calculation (typically by two orders of magnitude) compared with algorithms which are based on the use of the whole profile (*e.g.* Harris *et al.*(1994) and Andreev and Bruce (1998)).

The next step is to adjust one of the parameters and recalculate the χ^2. The new parameter value is accepted if a reduction in χ^2 is observed. If χ^2 increases, the likelihood of acceptance is

subject to a Boltzmann distribution. As a result of the Boltzmann condition, the chances of accepting an uphill move are higher at higher temperature:

if $\chi_i > \chi_j$: accept

if $\chi_i < \chi_j$: accept only if r < $e^{\left(\chi_i^2 - \chi_j^2/T\right)}$

where *i* and *j* correspond to the starting and new sets of variables respectively; *r* is a random number between 0 and 1.

The remaining parameters are adjusted in sequence a total of N_1 times, followed by an adjustment of step length of each of the parameters (such that the 50:50 accept:reject ratio of solutions is maintained). The algorithm then returns to the parameter adjusting step (step 3 in Figure 1.5) and the sequence is repeated N_2 times. Only then is the temperature reduced.

Unlike most SA algorithms, the temperature reduction is not performed at a constant rate and follows the Equation 1.4. As a consequence, when broad χ^2 fluctuations are observed, the CR is reduced allowing these regions to be comprehensively explored.

$$T_{next} = \frac{T_{current}}{1 + [T_{current} CR/3(\langle \chi^2 - \langle \chi^2 \rangle \rangle^2)^{1/2}]} \qquad \text{Equation 1.4}$$

where *CR* is the cooling rate and $(\langle \chi^2 - \langle \chi^2 \rangle \rangle^2)^{1/2}$ is the mean-square deviation of the χ^2.

The annealing terminates when $N_1 \times N_2 \times \text{DoF} \geq$ Max number of SA moves, or if a predefined profile χ^2 ratio has been reached. Normally, a simplex minimisation is applied at the end of the run (user configurable), but users may also manually invoke the simplex during the run to (possibly) speed up termination. It is important to remember that no single, finite, SA run is guaranteed to be able to locate the global minimum and so multiple SA runs are normally employed in order to improve the chances of success.

Although DASH has been mainly developed as a program aimed at crystal structure solution, rather than Rietveld refinement, a rigid-body refinement is also implemented and can be applied to any of the solutions obtained from a series of SA runs.

In summary, DASH provides “data to solution” functionality, which is effective and user-friendly, and it has been successfully used for the crystal structure determination of a large number of industrially important molecules.

1.5 The current limits of PXRD

A recent study of the data deposited in the CSD has shown that the organic structures solved from powder diffraction data constitute only approximately 0.5% of the total number of organic structures deposited (Cole *et al.*, 2014). Despite this, similar trends are observed, between the single-crystal and powder crystal structures deposited, for a number of factors (Figure 1.6-Figure 1.8). These include the relative increase of structures deposited since the 1990s and the mean structural complexity of the deposited crystal structures (both in terms of total number of atoms in the asymmetric unit and the DoF).

By way of example, there were 26 crystal structures determined from powder diffraction data deposited in the year 2000; in 2009, there were 79, a 3-fold increase. The non-powder CSD depositions doubled over same time period. Furthermore, the complexity of organic structures solved from powder diffraction data is now, on average, comparable to that of structures determined by single-crystal diffraction (Figure 1.6 - Figure 1.8).

The figures also show that despite some recent advances in the area of SDPD, there has been no significant increase in the mean structural complexity of solved structures since 2000. It is not unreasonable to conclude that the current limits of accessible structural complexity have been reached and as such, it is crucial to introduce new methods which will enhance the SDPD applicability. It is also possible that this 'flattening off' of structural complexity is merely a reflection of the compounds that are currently of interest to structural scientists.

A representative collection of crystal structures, solved with PXRD, is presented in Table 1.2. The examples have been selected on the basis of pharmaceutical and historical interest. For comparison the number of non-hydrogen atoms in the asymmetric unit and the DoF are listed, together with information on the methods and programs used. The corresponding molecular structures are shown in Figure 1.9.

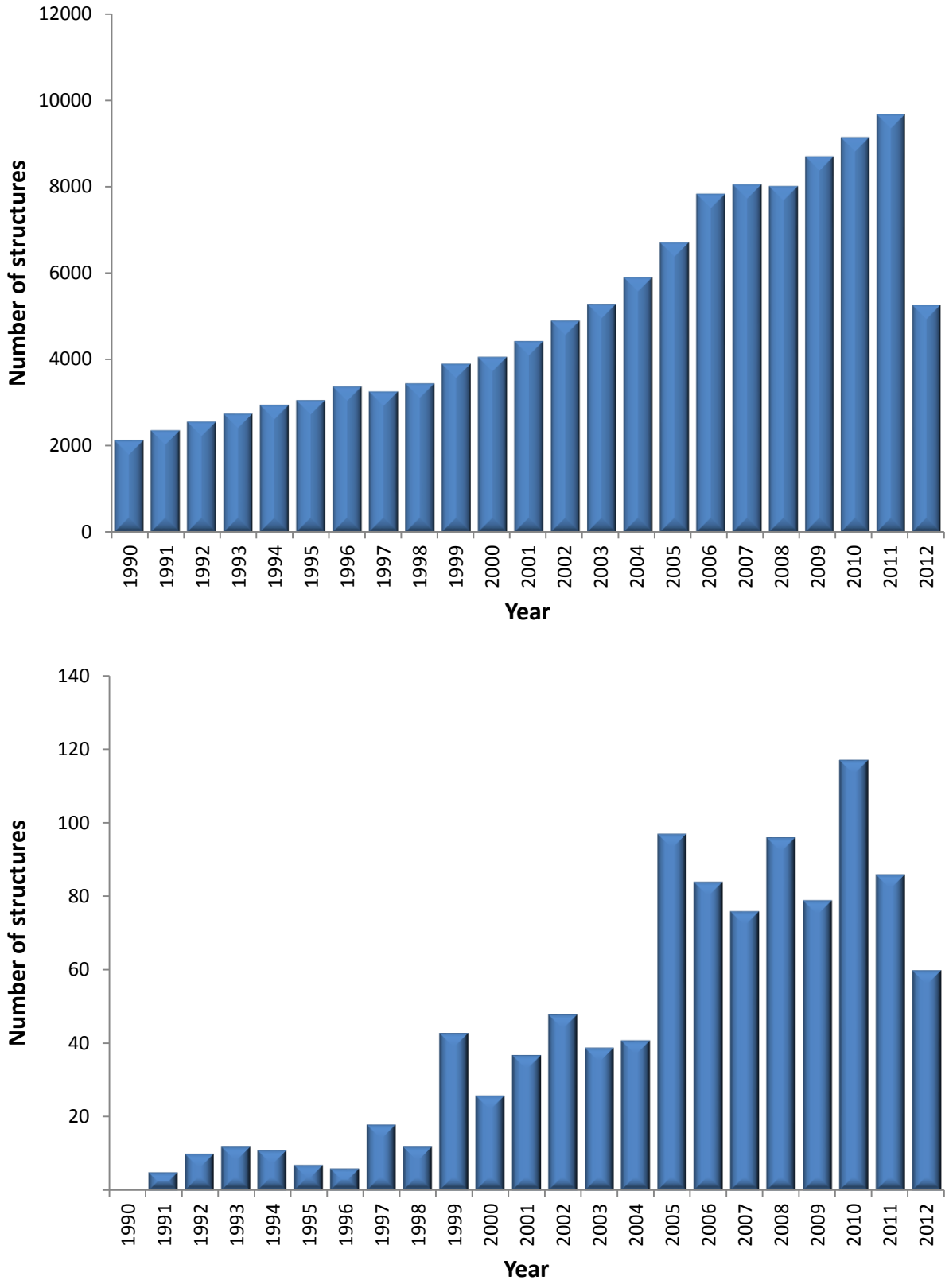


Figure 1.6. The number of molecular crystal structures deposited each year since 1990 in the CSD. The upper plot is derived from all structures whilst the lower plot is derived from only powder structures. Note that the plot is based on the November 2012 release of the CSD and, therefore, the final total for 2012 will be higher than the value plotted here.

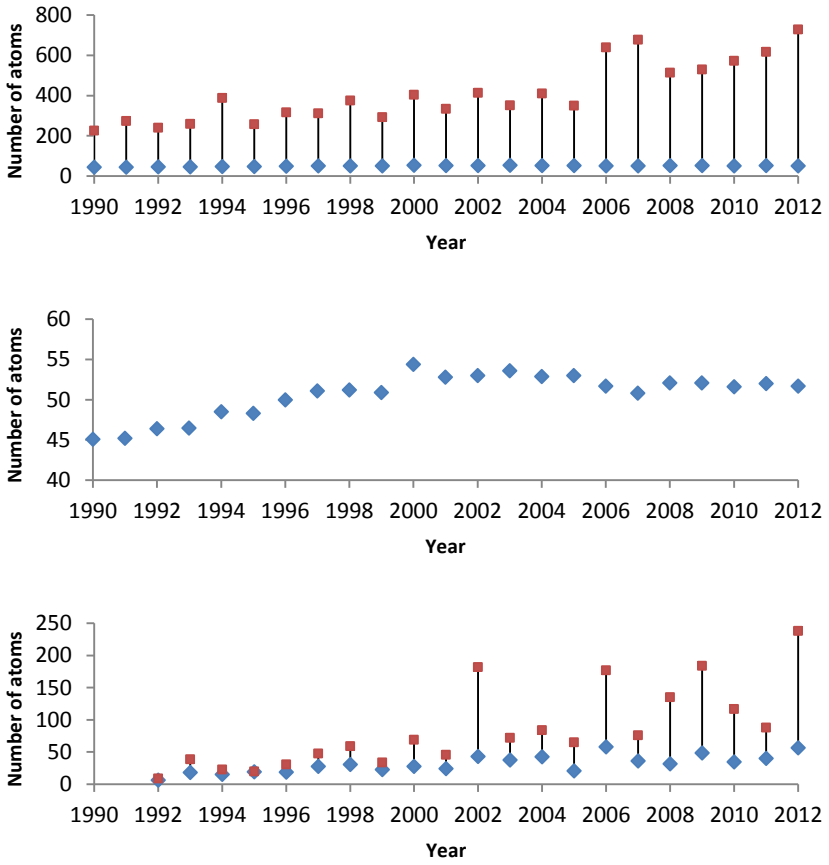


Figure 1.7. The mean (blue diamond) and maximum (red square) of the total number of atoms (including hydrogen) in the asymmetric unit of molecular crystal structures deposited each year since 1990 in the CSD. The upper plot is derived from all structures; the middle plot shows an expanded view of the mean number of atoms for all structures; the lower plot is derived from only powder structures.

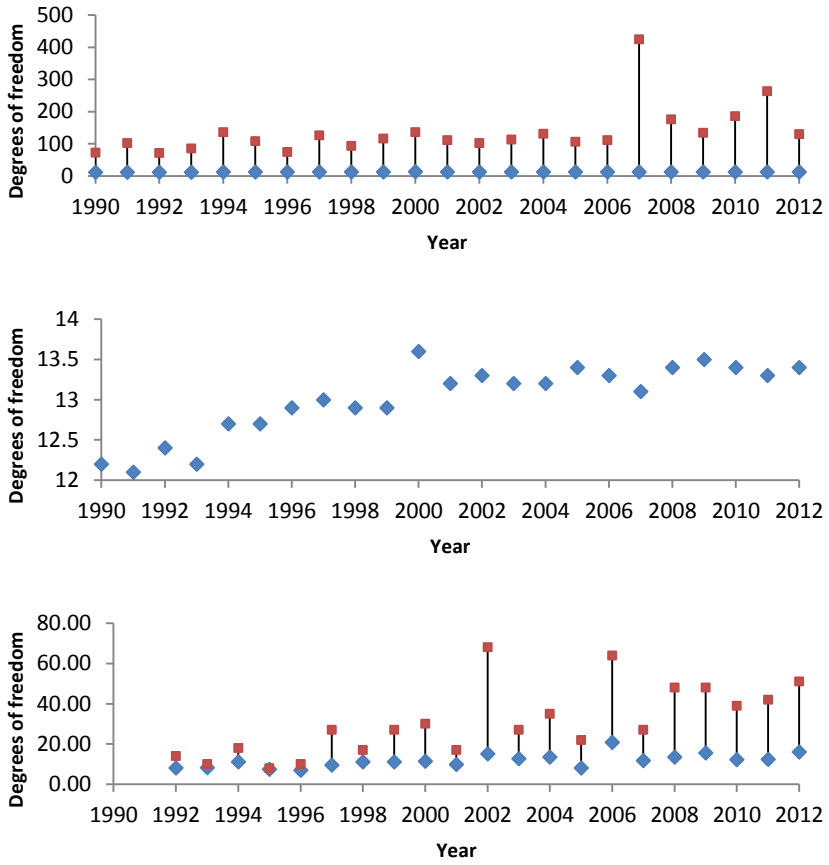


Figure 1.8.The mean (blue diamond) and maximum (red square) total number of degrees of freedom in the asymmetric unit of molecular crystal structures deposited each year since 1990 in the CSD. The upper plot is derived from all structures; the middle plot shows an expanded view of the mean number of degrees of freedom for all structures; the lower plot is derived from only powder structures.

Table 1.2. A summary of SDPD crystal structures, selected on the basis of pharmaceutical and historical interest. The corresponding molecular structures are shown in Figure 1.9. N_{atom} = total number of atoms in the asymmetric unit; N_{nonH} = total number on non-hydrogen atoms in the asymmetric unit; DoF = total number of degrees of freedom in the asymmetric unit. Key: CDE = cultural differential evolution; DM = direct methods; GA = genetic algorithm; GS = grid search; HBB–BC = hybrid big bang–big crunch; HMC = hybrid Monte Carlo; HY = hybrid methods; LE = Lamarckian evolution; LM = local minimization; MDM = modified direct methods; PM = Patterson methods; PS = particle swarm; PT = parallel tempering; SA = simulated annealing; SE = structure envelope

Name	2D	Z'	N_{atom}	N_{nonH}	DoF	Method	Software	Reference
Salicylic acid	a	1	16	10	7	GS	P-RISCON	(Masciocchi *et al.*, 1994)
Chlorothiazide	b	1	23	17	7	DM	MITHRIL94	(Shankland *et al.*, 1997)
Ibuprofen	c	1	33	15	10	GA	GAP	(Shankland *et al.*, 1998)
L-glutamic acid	d	1	19	10	10	LE	-	(Turner *et al.*, 2000)
Remacemide nitrate	e	1	45	24	18	SA	DASH	(Markvardsen *et al.*, 2002)
Tri-β-peptide	f	1	94	41	23	SE+SA	Safe	(Brenner *et al.*, 2002)
Capsaicin	g	1	49	22	15	HMC	-	(Markvardsen *et al.*, 2005)
Baicalein	h	1	30	20	7	CDE-GA	-	(Chong and Tremayne, 2006)
Famotidine	i	1	35	20	13	PM	EXPO	(Burla *et al.*, 2007)
Caffeine	j	5	120	70	30	SA	TOPAS	(Lehmann and Stowasser, 2007)
Captopril	k	1	29	14	10	MDM	EXPO	(Altomare *et al.*, 2007)
Chlorothiazide	b	1	23	17	7	DM	SHELX	(Fernandes *et al.*, 2008)
Cyheptamide	l	4	132	72	28	SA	DASH	(Florence *et al.*, 2008)
Tolbutamide	m	1	36	18	13	PS	PeckCryst	(Feng *et al.*, 2009)
Tolbutamide	m	1	36	18	13	GA	GEST	(Feng and Dong, 2007)
Capsaicin	g	1	49	22	15	LM	-	(Shankland *et al.*, 2010)
Nifedipine	n	2	86	50	24	SA	ReX	(Bortolotti *et al.*, 2011)
L-arginine	o	2	52	24	25	GA	EAGER	(Courvoisier *et al.*, 2012)
Amodiaquinium dichloride dihydrate	p	1	57	29	30	MDM	EXPO	(Altomare *et al.*, 2012)
Vorinostat	q	1	39	19	16	SA/PT	FOX	(Puigjaner *et al.*, 2012)
Amcinonide	r	2	142	72	20	SA	PSSP	(Pagola and Stephens, 2012)
Diphenhydramine hydrochloride	s	1	42	20	15	LM	TALP	(Vallcorba *et al.*, 2012)
Verapamil hydrochloride	t	1	73	34	22	HBB-BC	EXPO	(Altomare *et al.*, 2013a)
Zopiclone dihydrate	u	1	50	29	16	HY	EXPO	(Altomare *et al.*, 2013c)
Prilocaine	v	1	36	16	12	SA	PowderSolve	(Rietveld *et al.*, 2013)

.

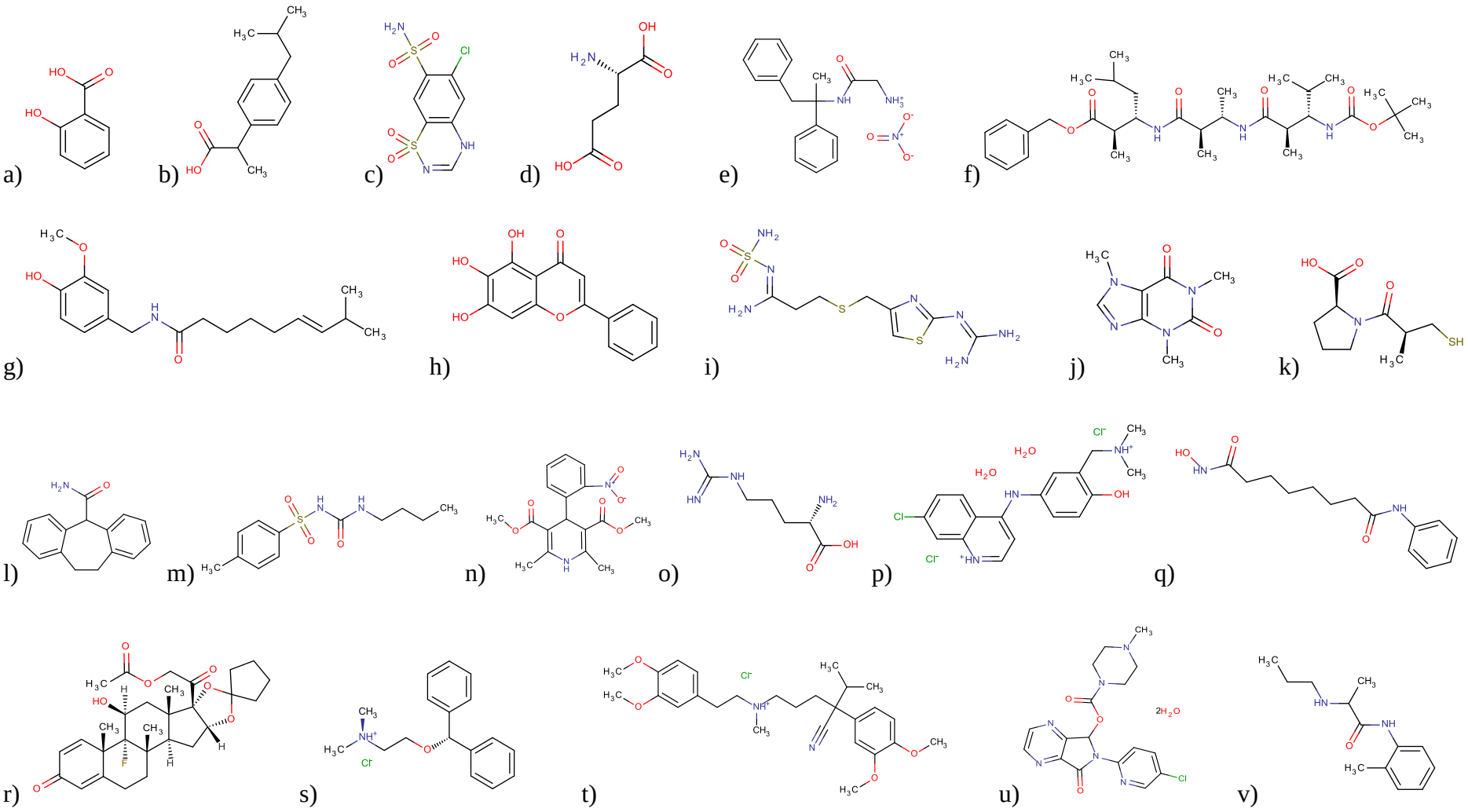


Figure 1.9. The molecular structures of the compounds listed in Table 1.2, obtained using MarvinSketch 'Name to Structure' (Chemaxon, 2011).

1.6 Enhancing SDPD

The enhancement of the current SDPD is the main focus of this thesis and whilst some of the aspects given below are discussed in detail in Chapters 4 and 6 they are briefly introduced here. It is also worth noting that the methods discussed are specifically targeted at enhancing GO methods.

1.6.1 Computational gains

Increasing the sheer compute power is an obvious way of tackling any computational problem. To reach the global minimum, GO algorithms for example require multiple runs which are independent of each other and can be performed in parallel over multiple cores. For example, carrying out 50 SA runs (each performing 10^7 moves) of verapamil hydrochloride takes approximately 24 hours. Utilising the 4 cores which a standard desktop PC has [MDASH, (Griffin *et al.*, 2009b)], the compute time can be reduced to 6 hours. Such calculation time gains can be particularly beneficial when faced with a very complex problem, which requires a large number of runs to maximise the chance of reaching the global minimum. Furthermore, with software such as GDASH (Griffin *et al.*, 2009a) and FOX.Grid (Rohlíček *et al.*, 2007) already in place (enabling the parallelisation/distribution of GO calculations), there is no reason why the possible gains should not be leveraged.

Of course the job processing time reduces with the increase of cores and the use of grid networks, such as the GRIDMP at STFC Rutherford Appleton laboratory and the UK national supercomputing services (HECToR), would result in gains orders of magnitude greater.

More recently, DASH was set-up to utilise cloud computing facilities [CDASH;(Spillman *et al.*, 2015)], providing an alternative 'pay-as-you-go' approach to distributed SDPD.

However, due to the exponential relationship between the problem hypersurface and the number of DoF, applying only sheer compute power has its limitations. Another approach would focus on a smarter use of methodologies (for example see 1.6.2); and furthermore, the combination of compute and methodological advances. There is no doubt however, that as the complexity of the structures being tackled from powder X-ray diffraction data increases, utilising parallelisation/distributed computing will become an integral part of the SDPD process.

1.6.2 Parameter tuning

GO algorithms have a number of user defined parameters which control their performance. As a result, finding appropriate algorithmic parameter values is a challenge for all software developers. Whilst it is commonly recognised as essential to achieve good software performance, in practice many algorithms only utilise generic values.

Very often, algorithm tuning is performed with well know mathematical functions such as the Griewank (1981), Rosenbrock (1960) and Rastrigin (1974).

Examples of parameter tuning of SA include the work of Park and Kim (1998) and Frausto-Solis *et al.* (2007), whilst GA tuning is comprehensively studied and reviewed by Eiben and co-workers (Eiben *et al.*, 1999; Eiben *et al.*, 2007; Eiben and Smit, 2011; Karafotias *et al.*, 2015) and others (Cao and Wu, 1999; Hoos, 2012; Huang *et al.*, 2014; Niaki *et al.*, 2014). Unsurprisingly, parameter tuning has been investigated more extensively for GA, due to the larger number of user definable parameters.

To the best of the author's knowledge, in the context of SDPD there are no reported examples of a SA parameter tuning, following the release of the SDPD software. DASH can be considered as a typical example, with its current default SA parameter values remaining unchanged since its first release. A partial reason for this is the general success DASH has delivered with small organic molecules. As such the SA parameter optimisation may be considered unnecessary. Due to the lack of further parameter investigation over the last decade however, it is indeed possible that there are SA parameter values which may improve the performance of DASH.

1.6.3 Prior conformational knowledge

As described above, GO methods already benefit from a vast amount of prior chemical information in the form of the molecular connectivity and well defined bond lengths and angles. Additionally, prior conformational knowledge can be utilised to confine the search space explored during the global optimisation, hence increasing the chances of crystal structure solution. A typical example can be given with the planar amide bond (R-C-N-R'), for which a value of 180^{o} can be confidently fixed, reducing the total DoF by one.

Solid-state NMR derived information has been found to be easily incorporated into SDPD; with examples varying from incorporation of Z' information (Harper *et al.*, 2005; Triponi *et al.*, 2014), to measurements of interatomic distances and interactions (Maruyoshi *et al.*, 2012;

Middleton *et al.*, 2002; Triponi *et al.*, 2014), and combining tensor information together with computational methods to exploit subtle differences of ^{35}Cl environments (Hamaed *et al.*, 2008).

The CSD system is another valuable tool for deriving prior conformational knowledge, the main advantage of which is that no additional experiments are required. Previously CSD derived conformational knowledge has been applied to individual examples, such as verapamil HCl (Florence *et al.*, 2005; Florence *et al.*, 2009), tetracaine hydrochloride (Nowell *et al.*, 2002) and verapamil HCl, famotidine and capsaicin (Cole *et al.*, 2014). Whilst these cases of exploiting additional conformational knowledge have shown it to be beneficial for SDPD, the method is not currently, routinely employed. This triggered an interest to carry out a comprehensive study of the benefits of prior conformational knowledge to the SDPD. A detailed discussion of the findings can be found in Chapter 6.

1.7 Aims and objectives

1.7.1 Aims

The aim of this project is to significantly extend the current limits of structural complexity that are accessible to global-optimisation-based SDPD methods. Specifically, it aims to enhance the performance of the DASH software package by employing methods that, whilst DASH-specific, are potentially transferrable to other GO methods.

1.7.2 Objectives

The key objectives are to:

1. Design and implement a protocol for the assembly of a comprehensive dataset of powder X-ray diffraction data, collected from molecular crystals.
2. Establish the current limits of applicability of DASH to crystal structure determination from powder X-ray diffraction data.
3. Investigate tuning of simulated annealing control parameters as a way of optimising the performance of DASH.
4. Investigate the effectiveness of different methods for the incorporation of prior conformational knowledge into DASH; specifically, the use of the Mogul-derived distributions and a novel conformer generator.

1.8 References

Allen FH, Johnson O, Shields GP, Smith BR and Towler M (2004) CIF applications. XV. enCIFer: a program for viewing, editing and visualizing CIFs. *J. Appl. Cryst.* **37**:335-338

Altomare A, Camalli M, Cuocci C, Giacovazzo C, Grazia A, Moliterni G and Rizzi R (2007) Direct methods and the solution of organic structures from powder data. *J. Appl. Cryst.* **40**:344-348

Altomare A, Corriero N, Cuocci C, Moliterni A and Rizzi R (2013a) The hybrid big bang-big crunch method for solving crystal structure from powder diffraction data. *J. Appl. Cryst.* **46**:779-787

Altomare A, Cuocci C, Giacovazzo C, Moliterni A and Rizzi R (2011) EXPO2011: A new package for powder crystallography. *Powder Diffr.* **26**:S2-S12

Altomare A, Cuocci C, Giacovazzo C, Moliterni A and Rizzi R (2012) COVMAP: a new algorithm for structure model optimization in the EXPO package. *J. Appl. Cryst.* **45**:789-797

Altomare A, Cuocci C, Giacovazzo C, Moliterni A, Rizzi R, Corriero N and Falcicchio A (2013b) EXPO2013: a kit of tools for phasing crystal structures from powder data. *J. Appl. Cryst.* **46**:1231-1235

Altomare A, Cuocci C, Moliterni A and Rizzi R (2013c) RAMM: a new random-model-based method for solving ab initio crystal structures using the EXPO package. *J. Appl. Cryst.* **46**:476-482

Andreev YG and Bruce PG (1998) Solving crystal structures of molecular solids without single crystals: a simulated annealing approach. *J. Chem. Soc.: Dalton Trans.*:4071-4080

Baerlocher C, McCusker LB and Palatinus L (2007) Charge flipping combined with histogram matching to solve complex crystal structures from powder diffraction data. *Z. Kristallogr.* **222**:47-53

Basso S, Besnard C, Wright JP, Margiolaki I, Fitch AN, Pattison P and Schiltz M (2010) Features of the secondary structure of a protein molecule from powder diffraction data. *Acta Cryst. Sect. D* **66**:756-761

Bortolotti M, Lonardelli I and Pepponi G (2011) Determination of the crystal structure of nifedipine form C by synchrotron powder diffraction. *Acta Cryst. Sect. B* **67**:357-364

Boultif A and Louer D (1991) Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method. *J. Appl. Cryst.* **24**:987-993

Boultif A and Louer D (2004) Powder pattern indexing with the dichotomy method. *J. Appl. Cryst.* **37**:724-731

Brenner S, McCusker LB and Baerlocher C (2002) The application of structure envelopes in structure determination from powder diffraction data. *J. Appl. Cryst.* **35**:243-252

Bricogne G (1984) Maximum-entropy and the foundations of direct methods. *Acta Cryst. Sect. A* **40**:410-445

Bricogne G (1991) A multisolution method of phase determination by combined maximization of entropy and likelihood .3. Extension to powder diffraction data. *Acta Cryst. Sect. A* **47**:803-829

Brittain HG (2001) X-ray diffraction III: Pharmaceutical applications of X-ray powder diffraction. *Spectr.* **16**:14

Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE and Orpen AG (2004) Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44**:2133-2144

Burla MC, Caliandro R, Carrozzini B, Cascarano GL, De Caro L, Giacovazzo C, Polidori G and Siliqi D (2007) The revenge of the Patterson methods. III. Ab initio phasing from powder diffraction data. *J. Appl. Cryst.* **40**:834-840

Cambridgesoft (1985-2015) ChemDraw, http://www.cambridgesoft.com, Oct 2015

Cao YJ and Wu QH (1999) Optimization of control parameters in genetic algorithms: a stochastic approach. *Int. J. Syst. Sci* **30**:551-559

Cerny R and Favre-Nicolin V (2007) Direct space methods of structure determination from powder diffraction: principles, guidelines and perspectives. *Z. Kristallogr.* **222**:105-113

Cerny V (1985) Thermodynamical approach to the traveling salesman problem - an efficient simulation algorithm. *J. Opt. Theory Appl.* **45**:41-51

ChemAxon (2011) Marvin 5.4.1.1, http:www.chemaxon.com, Oct 2015

Chieng N, Rades T and Aaltonen J (2011) An overview of recent studies on the analysis of pharmaceutical polymorphs. *J. Pharm. Biomed. Anal.* **55**:618-644

Chong SY and Tremayne M (2006) Combined optimization using cultural and differential evolution: application to crystal structure solution from powder diffraction data. *Chem. Comm.*:4078-4080

Clemons WK, Grundel DA and Jeffcoat DE (2004) Applying simulated annealing to the multidimensional assignment problem, in *Theory and Algorithms for Cooperative Systems* (Grundel D, Murphey R and Pardalos PM eds) pp 45-61

Coelho A (2003) TOPAS User Manual. Bruker AXS GmbH, Karlsruhe,.

Cole JC, Kabova EA and Shankland K (2014) Utilizing organic and organometallic structural data in powder diffraction. *Powder Diffr.* **29**:S19-S30

Courvoisier E, Williams PA, Lim GK, Hughes CE and Harris KDM (2012) The crystal structure of L-arginine. *Chem. Comm.* **48**:2761-2763

Datta S and Grant DJW (2004) Crystal structures of drugs: Advances in determination, prediction and engineering. *Nature Rev. Drug Discov.* **3**:42-57

David WIF and Shankland K (2008) Structure determination from powder diffraction data. *Acta Cryst. Sect. A* **64**:52-64

David WIF, Shankland K and Shankland N (1998) Routine determination of molecular crystal structures from powder diffraction data. *Chem. Comm.* 10.1039/a800855h:931-932

David WIF, Shankland K, van de Streek J, Pidcock E, Motherwell WDS and Cole JC (2006) DASH: a program for crystal structure determination from powder diffraction data. *J. Appl. Cryst.* **39**:910-915

Debets P (1968) The structures of uranyl chloride and its hydrates. *Acta Cryst. Sect. B* **24**:400-402

Dollase WA (1986) Correction of intensities for preferred orientation in powder diffractometry - application of the march model. *J. Appl. Cryst.* **19**:267-272

Eiben AE, Hinterding R and Michalewicz Z (1999) Parameter control in evolutionary algorithms. *Ieee Trans. Evol. Comput.* **3**:124-141

Eiben AE, Michalewicz Z, Schoenauer M and Smith JE (2007) Parameter control in evolutionary algorithms, in *Parameter Setting in Evolutionary Algorithms* (Lobo FG, Lima CF and Michalewicz Z eds) pp 19-46

Eiben AE and Smit SK (2011) Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm Evol. Comput.* **1**:19-31

Engel GE, Wilke S, Konig O, Harris KDM and Leusen FJJ (1999) PowderSolve - a complete package for crystal structure solution from powder diffraction patterns. *J. Appl. Cryst.* **32**:1169-1179

Favre-Nicolin V and Cerny R (2002) FOX, 'free objects for crystallography': a modular approach to ab initio structure determination from powder diffraction. *J. Appl. Cryst.* **35**:734-743

Feng ZJ and Dong C (2007) GEST: a program for structure determination from powder diffraction data using a genetic algorithm. *J. Appl. Cryst.* **40**:583-588

Feng ZJ, Dong C, Jia RR, Di Deng X, Cao SX and Zhang JC (2009) PeckCryst: a program for structure determination from powder diffraction data using a particle swarm optimization algorithm. *J. Appl. Cryst.* **42**:1189-1193

Fernandes P, Shankland K, David WIF, Markvardsen AJ, Florence AJ, Shankland N and Leech CK (2008) A differential thermal expansion approach to crystal structure determination from powder diffraction data. *J. Appl. Cryst.* **41**:1089-1094

Fernandes P, Shankland K, Florence AJ, Shankland N and Johnston A (2007) Solving molecular crystal structures from X-ray powder diffraction data: The challenges posed by γ-carbamazepine and chlorothiazide N,N,-dimethylformamide (1/2) solvate. *J. Pharm. Sci.* **96**:1192-1202

Florence AJ (2009) Approaches to High-Throughput Physical Form Screening and Discovery, in *Polymorphism in Pharmaceutical Solids, Second Edition* pp 139-184, CRC Press

Florence AJ, Shankland K, Gelbrich T, Hursthouse MB, Shankland N, Johnston A, Fernandes P and Leech CK (2008) A catemer-to-dimer structural transformation in cyheptamide. *CrystEngComm* **10**:26-28

Florence AJ, Shankland N, Shankland K, David WIF, Pidcock E, Xu XL, Johnston A, Kennedy AR, Cox PJ, Evans JSO, Steele G, Cosgrove SD and Frampton CS (2005) Solving molecular crystal structures from laboratory X-ray powder diffraction data with DASH: the state of the art and challenges. *J. Appl. Cryst.* **38**:249-259

Florence AJ, Taylor R, Shankland N and Shankland K (2009) Applications of the CSD to structure determination from powder data. *Abs. Pap. Am. Chem. Soc.* **237**

Frausto-Solis J, Alonso-Pecina F and Gonzalez-Segura C (2007) Analytically tuned parameters of simulated annealing for the Timetabling problem, in *Cimmacs '07: Proceedings of the 6th Wseas International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics* (Katehakis MN, Andina D and Mastorakis N eds) pp 18-23

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery Jr. JA, Peralta JE, Ogliaro F, Bearpark MJ, Heyd J, Brothers EN, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam NJ, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J and Fox DJ (2009) Gaussian 09. Gaussian, Inc., Wallingford, CT, USA.

Gilmore CJ and Bricogne G (1997) MICE computer program, in *Macromolecular Crystallography, Pt B* (Carter CW and Sweet RM eds) pp 65-78

Granovsky AA Firefly version 8, http://classic.chem.msu.su/gran/firefly/index.html, Oct 2015

Griewank AO (1981) Generalized descent for global optimization. *J. Opt. Theory Appl.* **34**:11-39

Griffin TAN, Shankland K, van de Streek J and Cole J (2009a) GDASH: a grid-enabled program for structure solution from powder diffraction data. *J. Appl. Cryst.* **42**:356-359

Griffin TAN, Shankland K, van de Streek J and Cole J (2009b) MDASH: a multi-core-enabled program for structure solution from powder diffraction data. *J. Appl. Cryst.* **42**:360-361

Hamaed H, Pawlowski JM, Cooper BFT, Fu R, Eichhorn SH and Schurko RW (2008) Application of solid-state ^{35}Cl NMR to the structural characterization of hydrochloride pharmaceuticals and their polymorphs. *J. Am. Chem. Soc.* **130**:11056-11065

Harper JK, Barich DH, Heider EM, Grant DM, Franke RR, Johnson JH, Zhang Y, Lee PL, Von Dreele RB, Scott B, Williams D and Ansell GB (2005) A combined solid-state NMR and X-ray powder diffraction study of a stable polymorph of Paclitaxel. *Cryst. Growth Des.* **5**:1737-1742

Harris KDM (2012) Powder Diffraction Crystallography of Molecular Solids, in *Advanced X-Ray Crystallography* (Rissanen K ed) pp 133-177

Harris KDM, Tremayne M, Lightfoot P and Bruce PG (1994) Crystal-structure determination from powder diffraction data by Monte-Carlo methods. *J. Am. Chem. Soc.* **116**:3543-3547

Hoos H (2012) Automated algorithm configuration and parameter tuning, in *Autonomous Search* (Hamadi Y, Monfroy E and Saubion F eds) pp 37-71, Springer Berlin Heidelberg

Huang Q, Zhang J, Zhang Q and Wei X (2014) Optimization of control parameters based on multi-objective genetic algorithms for spacecraft large angle attitude. *Mechaical, Electron. and Eng. Tech. (Icmeet 2014)* **538**:470-475

Huang XY, Miyazaki Y and Kajitani T (2008) High temperature thermoelectric properties of $Ca_{1-x}Bi_xMn_{1-y}V_yO_{3-delta}$ (0 <= x=y <= 0.08). *Solid State Commun.* **145**:132-136

IUCr (2014) *checkCIF*, http://checkcif.iucr.org/, Oct 2015

Ivanisevic I, McClurg RB and Schields PJ (2010) Uses of X-Ray Powder Diffraction In the Pharmaceutical Industry, in *Pharmaceutical Sciences Encyclopedia* (Gad SC ed), John Wiley & Sons, Inc.

Izumi F (2004) Beyond the ability of Rietveld analysis: MEM-based pattern fitting. *Solid State Ionics* **172**:1-6

Johnston JC, David WIF, Markvardsen AJ and Shankland K (2002) A hybrid Monte Carlo method for crystal structure determination from powder diffraction data. *Acta Cryst. Sect. A* **58**:441-447

Jung DS, Baerlocher C, McCusker LB, Yoshinari T and Seebach D (2014) Solving the structures of light-atom compounds with powder charge flipping. *J. Appl. Cryst.* **47**:1569-1576

Karafotias G, Hoogendoorn M and Eiben AE (2015) Parameter control in evolutionary algorithms: Trends and challenges. *Ieee Trans. Evol. Comput.* **19**:167-187

Kariuki BM, Calcagno P, Harris KDM, Philp D and Johnston RL (1999) Evolving opportunities in structure solution from powder diffraction data - Crystal structure determination of a molecular system with twelve variable torsion angles. *Angew. Chem. Int. Ed.* **38**:831-835

Kariuki BM, Serrano-Gonzalez H, Johnston RL and Harris KDM (1997) The application of a genetic algorithm for solving crystal structures from powder diffraction data. *Chem. Phys. Lett.* **280**:189-195

Kirkpatrick S, Gelatt CD and Vecchi MP (1983) Optimization by simulated annealing. *Science* **220**:671-680

Larson AC and Von Dreele RB (1994) General structure analysis system (GSAS). Laboratory LAN, LAUR 86-748, Los Alamos, California,

Le Bail A, Duroy H and Fourquet JL (1988) Ab-initio structure determination of $LiSbWO_6$ by X-ray powder diffraction. *Mater. Res. Bull.* **23**:447-452

Lehmann CW and Stowasser F (2007) The crystal structure of anhydrous beta-caffeine as determined from X-ray powder-diffraction data. *Chem. Eur. J.* **13**:2908-2911

Lemmerer A, Bernstein J, Griesser UJ, Kahlenberg V, Toebbens DM, Lapidus SH, Stephens P and Esterhuysen C (2011) A tale of two polymorphic pharmaceuticals: Pyrithyldione and Propyphenazone and their 1937 Co-Crystal Patent. *Chem. Eur. J.* **17**:13445-13460

Li H and Liu C (2013) Prediction of protein structures using a map-reduce hadoop framework based simulated annealing algorithm, in *2013 Ieee International Conference on Bioinformatics and Biomedicine* (Li GZ, Kim S, Hughes M, McLachlan G, Sun H, Hu X, Ressom H, Liu B and Liebman M eds)

Lim GK, Fujii K, Harris KDM and Apperley DC (2011) Structure Determination from Powder X-ray Diffraction Data of a New Polymorph of a High-Density Organic Hydrate Material, with an Assessment of Hydrogen-Bond Disorder by Rietveld Refinement. *Cryst. Growth Des.* **11**:5192-5199

Madsen IC and Hill RJ (1994) Collection and analysis of powder diffraction data with near-constant counting statistics. *J. Appl. Cryst.* **27**:385-392

Markvardsen AJ, David WIF and Shankland K (2002) A maximum-likelihood method for global-optimization-based structure determination from powder diffraction data. *Acta Cryst. Sect. A* **58**:316-326

Markvardsen AJ, Shankland K, David WIF and Didlick G (2005) Characterization of a hybrid Monte Carlo search algorithm for structure determination. *J. Appl. Cryst.* **38**:107-111

Markvardsen AJ, Shankland K, David WIF, Johnston JC, Ibberson RM, Tucker M, Nowell H and Griffin T (2008) ExtSym: a program to aid space-group determination from powder diffraction data. *J. Appl. Cryst.* **41**:1177-1181

Maruyoshi K, Iuga D, Antzutkin ON, Alhalaweh A, Velagad SP and Brown SP (2012) Identifying the intermolecular hydrogen-bonding supramolecular synthons in an indomethacin-nicotinamide cocrystal by solid-state NMR. *Chem. Comm.* **48**:10844-10846

Masciocchi N, Bianchi R, Cairati P, Mezza G, Pilati T and Sironi A (1994) P-RISCON - a real-space scavenger for crystal-structure determination from powder diffraction data. *J. Appl. Cryst.* **27**:426-429

Middleton DA, Peng X, Saunders D, Shankland K, David WIF and Markvardsen AJ (2002) Conformational analysis by solid-state NMR and its application to restrained structure determination from powder diffraction data. *Chem. Comm.*:1976-1977

Mohagheghian E, Bahadori A and James LA (2015) Carbon dioxide compressibility factor determination using a robust intelligent method. *J. Supercrit. Fluids* **101**:140-149

Morissette SL, Almarsson Ö, Peterson ML, Remenar JF, Read MJ, Lemmo AV, Ellis S, Cima MJ and Gardner CR (2004) High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Adv. Drug Deliv. Rev.* **56**:275-300

Moschakis IA and Karatza HD (2015) Towards scheduling for Internet-of-Things applications on clouds: a simulated annealing approach. *Concurrency and Comput.: Pract. Exper.* **27**:1886-1899

Neumann M (2003) X-Cell: a novel indexing algorithm for routine tasks and difficult cases. *J. Appl. Cryst.* **36**:356-365

Niaki STA, Gazaneh FM and Toosheghanian M (2014) A Parameter-Tuned Genetic Algorithm for Economic-Statistical Design of Variable Sampling Interval X-Bar Control Charts for Non-Normal Correlated Samples. *Commun. Stat. Simul. Comput.* **43**:1212-1240

Nishibori E, Ogura T, Aoyagi S and Sakata M (2008) Ab initio structure determination of a pharmaceutical compound, prednisolone succinate, from synchrotron powder data by combination of a genetic algorithm and the maximum entropy method. *J. Appl. Cryst.* **41**:292-301

Noguchi S, Miura K, Fujiki S, Iwao Y and Itai S (2012) Clarithromycin form I determined by synchrotron X-ray powder diffraction. *Acta Cryst. Sect. C* **68**:O41-O44

Nowell H, Attfield JP, Cole JC, Cox PJ, Shankland K, Maginn SJ and Motherwell WDS (2002) Structure solution and refinement of tetracaine hydrochloride from X-ray powder diffraction data. *New J. Chem.* **26**:469-472

Oszlanyi G and Suto A (2004) Ab initio structure solution by charge flipping. *Acta Cryst. Sect. A* **60**:134-141

Oszlanyi G and Suto A (2008) The charge flipping algorithm. *Acta Cryst. Sect. A* **64**:123-134

Pagola S and Stephens PW (2010) PSSP, a computer program for the crystal structure solution of molecular materials from X-ray powder diffraction data. *J. Appl. Cryst.* **43**:370-376

Pagola S and Stephens PW (2012) Structural study of an unusually large molecular solid from powder diffraction: the sequential unravelling of hydrogen bonding and van der Waals interactions contributing to the Z '= 2 crystal packing of amcinonide. *CrystEngComm* **14**:5349-5354

Paolo G, Stefano B, Nicola B, Matteo C, Roberto C, Carlo C, Davide C, Guido LC, Matteo C, Ismaila D, Andrea Dal C, Stefano de G, Stefano F, Guido F, Ralph G, Uwe G, Christos G, Anton K, Michele L, Layla M-S, Nicola M, Francesco M, Riccardo M, Stefano P, Alfredo P, Lorenzo P, Carlo S, Sandro S, Gabriele S, Ari PS, Alexander S, Paolo U and Renata MW (2009) QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **21**:395502

Park MW and Kim YD (1998) A systematic procedure for setting parameters in simulated annealing algorithms. *Comput. Operat. Res.* **25**:207-217

Paszkowicz W (2009) Genetic algorithms, a nature-inspired tool: Survey of applications in materials science and related fields. *Mater. Manuf. Processes* **24**:174-197

Patterson A (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev. A: At. Mol. Opt. Phys.* **46**:372

Pawley G (1981) Unit-cell refinement from powder diffraction scans. *J. Appl. Cryst.* **14**:357-361

Petricek V, Dusek M and Palatinus L (2014) Crystallographic computing system JANA2006: General features. *Z. Kristallogr.* **229**:345-352

Pinar AB, Gomez-Hortigueela L, McCusker LB and Perez-Pariente J (2011) Synthesis of Zn-containing microporous aluminophosphate with the STA-1 structure. *Dalton Trans.* **40**:8125-8131

Puigjaner C, Barbas R, Portell A, Valverde I, Vila X, Alcobe X, Font-Bardia M and Prohens R (2012) A cocrystal is the key intermediates for the production of a new polymorph of Vorinostat. *CrystEngComm* **14**:362-365

Putz H, Schon JC and Jansen M (1999) Combined method for ab initio structure solution from powder diffraction data. *J. Appl. Cryst.* **32**:864-870

Radu C and Vintan L (2013) Domain-knowledge optimized simulated annealing for network-on-chip application mapping, in *Advances in Intelligent Control Systems and Computer Science* (Dumitrache I ed) pp 473-487

Rahimian F, Payberah AH, Girdzijauskas S, Jelasity M and Haridi S (2015) A distributed algorithm for large-scale graph partitioning. *ACM Trans. Auton. Adapt. Syst.* **10**:1-24

Randall C, Rocco W and Ricou P (2010) XRD in Pharmaceutical Analysis: A Versatile Tool for Problem-Solving. *Am. Pharm. Rev.* **13**:52-59

Rastrigin LA (1974) Extremal control systems, in *Theoretical Foundations of Engineering Cybernetics Series*, Moscow

Renzi C, Leali F, Cavazzuti M and Andrisano AO (2014) A review on artificial intelligence applications to the optimal design of dedicated and reconfigurable manufacturing systems. *Int. J. Adv. Manufact. Tech.* **72**:403-418

Rietveld H (1969) A profile refinement method for nuclear and magnetic structures. *J. Appl. Cryst.* **2**:65-71

Rietveld IB, Perrin M-A, Toscani S, Barrio M, Nicolai B, Tamarit J-L and Ceolin R (2013) Liquid-Liquid Miscibility Gaps in Drug-Water Binary Systems: Crystal Structure and Thermodynamic

Properties of Prilocaine and the Temperature-Composition Phase Diagram of the Prilocaine-Water System. *Mol. Pharmaceut.* **10**:1332-1339
Rius J (1999) XLENS, a direct methods program based on the modulus sum function: Its application to powder data. *Powder Diffr.* **14**:267-273
Rius J (2011) Patterson-function direct methods for structure determination of organic compounds from powder diffraction data. XVI. *Acta Cryst. Sect. A* **67**:63-67
Rius J, Labrador A, Crespi A, Frontera C, Vallcorba O and Carles Melgarejo J (2011) Capabilities of through-the-substrate microdiffraction: application of Patterson-function direct methods to synchrotron data from polished thin sections. *J. Synch. Rad.* **18**:891-898
Rodriguez-Carvajal J (1993) FullProf, https://www.ill.eu/sites/fullprof/index.html, Oct 2015
Rohlíček J, Hušák M and Favre-Nicolin V (2007) Fox.Grid, http://fox.vincefn.net/Manual/Fox.Grid., Oct 2015
Rosenbrock HH (1960) An automatic method for finding the greatest or least value of a function. *The Comput. J.* **3**:175-184
Shankland K and David WIF (2002) Global optimization strategies, in *Structure Determination from Powder Diffraction Data* (David WIF, Shankland K, McCusker LB and Baerlocher C eds) pp 252-285, Oxford University Press, United States
Shankland K, David WIF, Csoka T and McBride L (1998) Structure solution of Ibuprofen from powder diffraction data by the application of a genetic algorithm combined with prior conformational analysis. *Int. J. Pharm.* **165**:117-126
Shankland K, David WIF, McCusker LB and Baerlocher C eds (2002) Structure determination from powder diffraction data, Oxford University Press, United States
Shankland K, David WIF and Sivia DS (1997) Routine ab initio structure determination of chlorothiazide by X-ray powder diffraction using optimised data collection and analysis strategies. *J. Mater. Chem.* **7**:569-572
Shankland K, Markvardsen AJ, Rowlatt C, Shankland N and David WIF (2010) A benchmark method for global optimization problems in structure determination from powder diffraction data. *J. Appl. Cryst.* **43**:401-406
Sitepu H, O'Connor BH and Li D (2005) Comparative evaluation of the March and generalized spherical harmonic preferred orientation models using X-ray diffraction data for molybdite and calcite powders. *J. Appl. Cryst.* **38**:158-167
Smith WE, Barrett HH and Paxman RG (1983) Reconstruction of objects from coded images by simulated annealing. *Opt. Lett.* **8**:199-201
Spek A (2003) Single-crystal structure validation with the program PLATON. *J. Appl. Cryst.* **36**:7-13
Spillman MJ, Shankland K, Williams AC and Cole JC (2015) CDASH: a cloud-enabled program for structure solution from powder diffraction data. *J. Appl. Cryst.* **48**:2033-2039
Stewart JJP (1990) Pecial issue - MOPAC - a semiempirical molecular-orbital program. *J. Comput. Aided Mol. Des.* **4**:1-45
Sun H-C and Huang Y-C (2012) Review of simulated annealing-based techniques for power system planning. *Int. Rev. Electr. Eng-Iree* **7**:5667-5677
Tanley SWM, Schreurs AMM, Kroon-Batenburg LMJ and Helliwell JR (2012) Room-temperature X-ray diffraction studies of cisplatin and carboplatin binding to His15 of HEWL after prolonged chemical exposure. *Acta Cryst. Sect. F* **68**:1300-1306
Treacy MMJ, Newsam JM and Deem MW (1989) Diffraction from zeolites containing planar faults, in *Characterization of the Structure and Chemistry of Defects in Materials* (Larson BC, Ruhle M and Seidman DN eds) pp 497-502, Cambridge University Press, New York, USA
Tremayne M (2004) The impact of powder diffraction on the structural characterization of organic crystalline materials. *Philos. Trans. R. Soc. London, Ser. A* **362**:2691-2707
Triponi C, Kacso I, Miclaus M, Filip X, Bratu I and Filip C (2014) Molecular structure elucidation of a new anhydrous polymorph of Acyclovir: Experimental and computational approach. *Rev. Chim.* **65**:657-663
Turner GW, Tedesco E, Harris KDM, Johnston RL and Kariuki BM (2000) Implementation of Lamarckian concepts in a Genetic Algorithm for structure solution from powder diffraction data. *Chem. Phys. Lett.* **321**:183-190

Vallcorba O, Rius J, Frontera C and Miravitlles C (2012) TALP: a multisolution direct-space strategy for solving molecular crystals from powder diffraction data based on restrained least squares. *J. Appl. Cryst.* **45**:1270-1277

Visser J (1969) A fully automatic program for finding the unit cell from powder data. *J. Appl. Cryst.* **2**:89-95

Werner P-E, Eriksson L and Westdahl M (1985) TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Cryst.* **18**:367-370

Xie D, Baerlocher C and McCusker LB (2011a) Using phases retrieved from two-dimensional projections to facilitate structure solution from X-ray powder diffraction data. *J. Appl. Cryst.* **44**:1023-1032

Xie D, McCusker LB and Baerlocher C (2011b) Structure of the Borosilicate Zeolite Catalyst SSZ-82 Solved Using 2D-XPD Charge Flipping. *J. Am. Chem. Soc.* **133**:20604-20610

Ye G and Rui X (2013) An improved simulated annealing andgenetic algorithm for TSP. *2013 5th Ieee Int. Conf. on Broadband Network & Multimedia Tech. (Ic-Bnmt)*:6-9

Yusup N, Zain AM and Hashim SZM (2012) Evolutionary techniques in optimizing machining parameters: Review and recent applications (2007-2011). *Expert Syst. Applic.* **39**:9909-9927

Zachariasen WH and Ellinger FH (1963) The crystal structure of beta plutonium metal. *Acta Cryst.* **16**:369-375

2 Materials and methods

This chapter summarises the main materials (compounds, crystal structures, associated PXRD data sets and computers) and methods (crystallographic and statistical) used throughout this work. Methods related to individual experiments and the experiments themselves are discussed in the relevant chapters.

2.1 Materials

2.1.1 Data sets for previously solved crystal structures

Powder diffraction data associated with one hundred and one, previously solved, crystal structures were assembled and divided into two sets: training (A1-A40) and test (B1-B61). A list of the structures, the codes by which they are referred to in this thesis, their CSD codes and their associated references is given in Table 2.1, whilst their corresponding molecular structures can be found in Figure 2.1. Further details on the selection and division criteria for these data sets are given in Chapter 3.

2.1.2 Laboratory X-ray data collection

Powder X-ray diffraction data for ritonavir (Sigma Aldrich, product code 91114, batch number 094M4709V) and lisinopril dihydrate (Sigma Aldrich, product code L0702000, batch number 2.0) were collected on a Bruker D8 Advance (Cu $K_{\alpha 1}$, λ= 1.54056 Å) diffractometer operating in capillary transmission mode. The diffractometer was equipped with a LynxEye detector. Monochromatic Cu $K\alpha_1$ is achieved with the use of a curved Johansson type primary monochromator. Furthermore, an 8 mm detector aperture slit and a metal knife edge collimator were used to minimise air scattering.

Both samples were used as received from Sigma-Aldrich, and the data collection was carried out at room temperature (*ca*. 293 K). Further details of the collection parameters are given in Table 2.2.

Table 2.1 Compound names and corresponding CSD reference codes of the 101 previously-solved crystal structures, together with the code names used throughout this thesis.

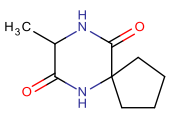
Code	Compound Name	CSD refcode	Reference
A1	Alaptide	KUTBEH	(Rohlicek *et al.*, 2010)
A2	Hydrochlorothiazide	HCSBTZ	(Dupont and Dideberg, 1972)
A3	Dapsone	DAPSUO10	(Alleaume, 1967)
A4	2-(Phenylsulfonyl)acetamide	Not deposited	(Florence *et al.*, 2005)
A5	Captopril	MCPRPL	(Fujinaga and James, 1980)
A6	Methyl 4-[(4-aminophenyl)ethynyl]-benzoate	Not deposited	(Florence *et al.*, 2005)
A7	Zopiclone	CUHNEY10	(Borea *et al.*, 1987)
A8	2-(4-Hydroxy-2-oxo-2,3-dihydro-1,3-benzothiazol-7-yl) ethylammonium chloride	BIFRAK	(Florence *et al.*, 2005)
A9	Salbutamol	BHHPHE	(Beale and Stephens, 1972)
A10	Dopamine hydrobromide	QQQAEJ01	(Shankland *et al.*, 1996)
A11	Chlorpropamide	BEDMIG	(Koo *et al.*, 1980)
A12	Creatine monohydrate	CREATH03	(Kato *et al.*, 1979)
A13	2,5-dioxopyrrolidin-1-yl 2-(benzoylsulfanyl) acetate	OQUPOG	(Rukiah and Al-Ktaifani, 2011)
A14	α-Lactose monohydrate	LACTOS10	(Fries *et al.*, 1971)
A15	Promazine hydrochloride	PROMZC01	(David *et al.*, 1998)
A16	Tolbutamide	ZZZPUS02	(Donaldson *et al.*, 1981)
A17	Carbamazepine dihydrate	FEFNOT01	(Florence *et al.*, 2005)
A18	Pigment orange 36 (PO 36)	HOYVOH	(van de Streek *et al.*, 2009)
A19	(4'-(2-(p-Tosylamino)benzylideneamino)-2,3-benzo-15-crown-5)-isothiocyanato-lithium	RIFVEI	(Dorokhov *et al.*, 2007)
A20	Famotidine	FOGVIG03	(Florence *et al.*, 2003)
A21	Sotalol hydrochloride	SOTALC	(Gadret *et al.*, 1976)
A22	Glipizide	SAXFED	(Burley, 2005)
A23	Diltiazem hydrochloride	CEYHUJ01	(Kojicprodic *et al.*, 1984)
A24	Zopiclone dihydrate	UCUVET	(Shankland *et al.*, 2001)
A25	Capsaicin	FABVAF01	(David *et al.*, 1998)
A26	Pigment yellow (PY 181 polymorph β)	GITWUC	(van de Streek *et al.*, 2009)
A27	Clarithromycin monohydrate	LAQSON	(Noguchi *et al.*, 2012a)
A28	Sodium 4-[(E)-(4-hydroxyphenyl)diazenyl] benzene sulfonate dihydrate	YAYWUQ	(Kennedy *et al.*, 2001)
A29	Indomethacin:nicotinamide 1:1	SESKUY	(Majumder *et al.*, 2013)
A30	Carbamazepine:indomethacin 1:1	LEZKEI	(Majumder *et al.*, 2013)
A31	2-[3-(2-Phenylethoxy)propyl sulfonyl] ethyl benzoate	BIFREO	(Florence *et al.*, 2005)
A32	S-Ibuprofen	JEKNOC10	(Freer *et al.*, 1993)
A33	Ampicilline trihydrate	AMPCIH01	(Burley *et al.*, 2006)
A34	Verapamil hydrochloride	CURHOM	(Carpy *et al.*, 1985)
A35	Amodiaquinium dichloride dihydrate	SENJIF	(Llinas *et al.*, 2006)
A36	Nifedipine (polymorph C)	BICCIZ01	(Bortolotti *et al.*, 2011)
A37	N-(2-(4-Hydroxy-2-oxo-2,3-dihydro-1,3-benzothiazol-7-yl)ethyl)-3-(2-(2-naphthalen-1-ylethoxy) ethylsulfonyl) propylaminium benzoate	PAHFIO	(Johnston *et al.*, 2004)
A38	Carbamazepine (polymorph γ)	CBMZPN13	(Fernandes *et al.*, 2007c)

Code	Compound Name	CSD refcode	Reference
A39	Cyheptamide	TEVSOD01	(Florence *et al.*, 2008)
A40	Ornidazole	NETRUZ	(Shin *et al.*, 1995)
B1	Tetraformaltrisazine	UDALIV	(Albov *et al.*, 2006)
B2	Decalin	POVZUW	(Eibl *et al.*, 2009)
B3	Pigment violet	QAMQOL	(Schmidt *et al.*, 2005)
B4	N,N'-Bis[1-pyridin-4-yl-meth-(E)-ylidene]hydrazine	LIZCUS	(Shanmuga Sundara Raj *et al.*, 2000)
B5	β- Phenazepam	BCHBZP01	(Sergeev *et al.*, 2010)
B6	2-Mercaptobenzoic acid	ZZZLWW01	(Steiner, 2000)
B7	Carbamazepine (polymorph β)	CBMZPN10	(Himes *et al.*, 1981)
B8	Hydroflumethiazide	EWUHAF	(Florence *et al.*, 2003)
B9	Paracetamol (polymorph I)	HXACAN07	(Nichols and Frampton, 1998)
B10	Paracetamol (polymorph II)	HXACAN08	(Nichols and Frampton, 1998)
B11	Phenylacetic acid	ZZZMLY01	(Hodgson and Asplund, 1991)
B12	5-anilinomethylene-2,2-dimethyl-1,3-dioxane-4,6-dione	MENMOI01	(Smrcok *et al.*, 2007)
B13	2,2,2-Trifluoro-N-(1a,2,7,7a-tetrahydronaphtho[2,3-b]oxiren-3-yl) acetamide	FAFQAG	(Rukiah and Assaad, 2010)
B14	Ethyl 1',2',3',4',4a',5',6',7'-octahydrodispiro[cyclohexane-1,2'-quinazoline-4',1''-cyclohexane]-8'-carbodithioate	RUJSOF	(Avila *et al.*, 2009)
B15	5-amino-3-[4-(3-methoxyphenyl)piperazin-1-yl]-1,2,3,4-tetrahydronaphthalen-2-ol	CALJOQ	(Assaad and Rukiah, 2011)
B16	(Z)-3-Methyl-N-(7-nitroacridin-3-yl)-2,3-dihydro-1,3-benzothiazol-2-imine	CALDOK	(Vallcorba *et al.*, 2011)
B17	trans-Dichlorobis(triphenylphosphine)nickel(II)	CLTPNI03	(Brammer and Stevens, 1989)
B18	Pamoic acid	DEGDAV	(Haynes *et al.*, 2006)
B19	4-(4’ -Dimethylaminostyryl)pyridine N-oxide	IJEKAJ	(Ivashevskaja *et al.*, 2003)
B20	2-(Benzoylsulfanyl)acetic acid	OQUPIA	(Rukiah and Al-Ktaifani, 2011)
B21	bis(4'-(2-(p-Tosylamino)benzylideneamino)-2,3-benzo-15-crown-5-N,N',O)-copper(ii)	RIFVAE	(Dorokhov *et al.*, 2007)
B22	trans-Di-isothiocyanato-bis(triphenylphosphine)-nickel	GEBZUI	(Bamgboye and Sowerby, 1986)
B23	Methyl 4-[4-(dimethylamino)phenyl]ethynyl benzoate	Not Deposited	(Marder, 2004)
B24	cis-Thiothixene	THTHXN01	(David *et al.*, 1998)
B25	Tetracycline hydrochloride	XAYCAB	(Clegg and Teat, 2000)
B26	Ezetimibe anhydrate	QUWYIR	(Bruning *et al.*, 2010)
B27	4-(Phenyldiazenyl)naphthalen-1-amine hydrochloride	QIJCAN	(Yatsenko *et al.*, 2001)
B28	3-azabicyclo[3.3.1]nonane-2,4-dione (form 2)	BOQQUT01	(Hulme *et al.*, 2006)
B29	1,4-Bis(2-phenethyloxyethanesulfonyl) piperazine	BIFRIS	(Florence *et al.*, 2005)
B30	3,5-Bis[(N,N-dimethylamino)methyl-eneamino]-1-methyl-4-nitropyrazole	WOCVUF	(Chernyshev *et al.*, 2000)
B31	Telmisartan (polymorph A)	XUYHOO01	(Dinnebier *et al.*, 2000)
B32	Telmisartan (polymorph B)	XUYHOO	(Dinnebier *et al.*, 2000)
B33	Clomipramine hydrochloride	CIMPRA	(Post and Horn, 1977)
B34	Clarithromycin (polymorph I)	NAVSUY02	(Noguchi *et al.*, 2012b)
B35	Pigment orange 62(PO 62)	HOYVUN	(van de Streek *et al.*, 2009)
B36	Pigment yellow (PY 151)	HOYWAU	(van de Streek *et al.*, 2009)

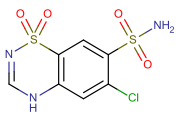
Code	Compound Name	CSD refcode	Reference
B37	Pigment yellow (PY 154 polymorph α)	HOYWEY	(van de Streek *et al.*, 2009)
B38	Pigment yellow 194 (PY 194)	HOYWIC	(van de Streek *et al.*, 2009)
B39	2,4-dinitro-N-phenyl-6-(phenylazo)-benzamide	IHESUJ	(Chernyshev *et al.*, 2002)
B40	N-methyl-2,4-dinitro-N-phenyl-6-(phenylazo)benzamide	IHETEU	(Chernyshev *et al.*, 2002)
B41	chlorothiazide N,N-dimethylformamide solvate	WEJHAV	(Fernandes *et al.*, 2006)
B42	Trihexyphenidyl hydrochloride	KUZDIT	(Maccaroni *et al.*, 2010)
B43	N-(2-methoxyphenyl)-2-(2-methoxyphenylazo)-4,6-dinitrobenzamide	IHETAQ	(Chernyshev *et al.*, 2002)
B44	Nimustine hydrochloride	WAWZAX	(Beko *et al.*, 2012)
B45	(R)-1-phenylethylammonium (R)-2-phenylbutyrate (polymorph II)	PBUPEA01	(Fernandes *et al.*, 2007a)
B46	(R)-1-phenylethylammonium (R)-2-phenylbutyrate (polymorph III)	PBUPEA02	(Fernandes *et al.*, 2007b)
B47	Tetracaine hydrochloride	XISVOK	(Nowell *et al.*, 2002)
B48	α/β-lactose	LAKKEO	(Lefebvre *et al.*, 2005)
B49	N-(6-Phenylhexanoyl)glycyltryptophanamide	FEFNOV	(Bushmarinov *et al.*, 2012)
B50	Pigment yellow 183 (PY183 polymorph α)	HOMMEC01	(Ivashevskaya *et al.*, 2009)
B51	Pigment yellow 191 (PY191 polymorph α)	HOMMIG01	(Ivashevskaya *et al.*, 2009)
B52	Pigment yellow 191 (PY191 polymorph β)	HOMMOM01	(Ivashevskaya *et al.*, 2009)
B53	Lisinopril dihydrate	GERWUX01	(Sorrenti *et al.*, 2013)
B54	Prednisolone succinate	KIXDEB01	(Nishibori *et al.*, 2008)
B55	Cytenamide (polymorph II)	TEVSOD01	(Florence *et al.*, 2008)
B56	Carvedilol dihydrogen phosphate propan-2-ol solvate	PUJTOE	(Chernyshev *et al.*, 2010)
B57	Ritonavir	YIGPIO01	(Bauer *et al.*, 2001)
B58	Crystal Violet Anhydrous	Not Deposited	Shankland, Private communication
B59	d-sorbitol	GLUCIT03	(Rukiah *et al.*, 2004)
B60	Chlorothiazide N,N-dimethylformamide solvate	NILSEH	(Fernandes *et al.*, 2007c)
B61	1,2,3,-tris(nonadecanoyl)glycerol (polymorph β)	MEZNAG	(Helmholdt *et al.*, 2002)

Table 2.2 Summary of the PXRD data collection parameters used for Ritonavir and Lisinopril dihydrate. VCT refers to the use of the variable count time scheme.

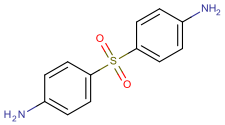
Compound	Range (˚2θ)	Step size (˚2θ)	VCT ranges (˚2θ)	Time per step (seconds)	Total collection (hours)	Capillary diameter (mm)
Lisinopril dihydrate	3-70.02	0.017	3.00-25.34 25.340-47.68 47.68- 70.02	9 18 28	20	0.5
Ritonavir	4-70.43	0.017	4.00-23.00 23.00- 41.99 41.99 – 60.99 60.99-70.43	6 8 16 32	15	0.7



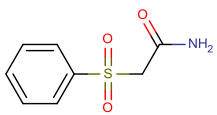
A1



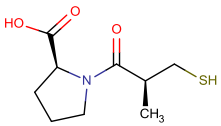
A2



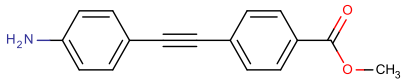
A3



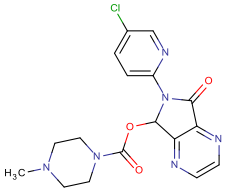
A4



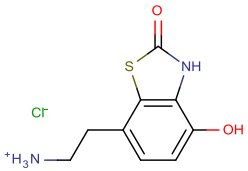
A5



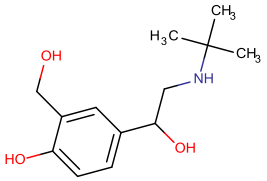
A6



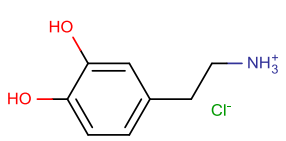
A7



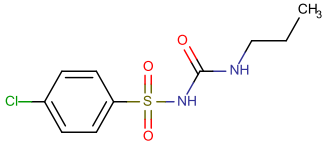
A8



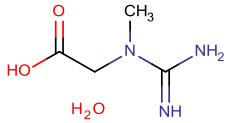
A9



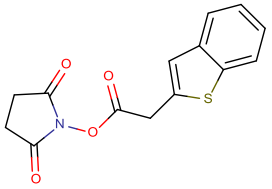
A10



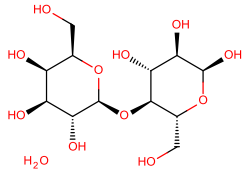
A11



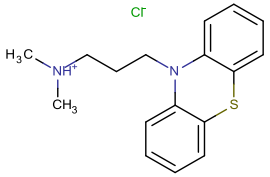
A12



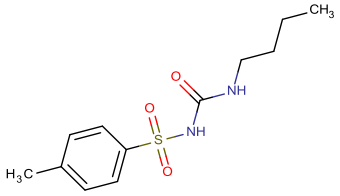
A13



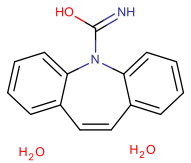
A14



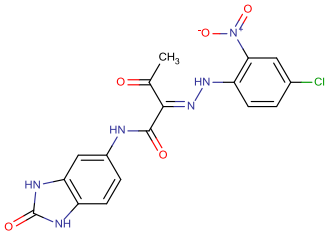
A15



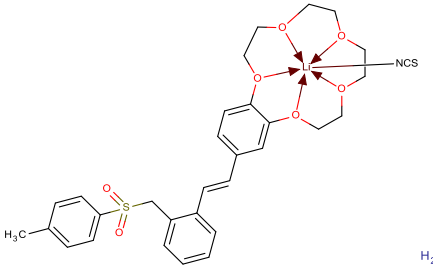
A16



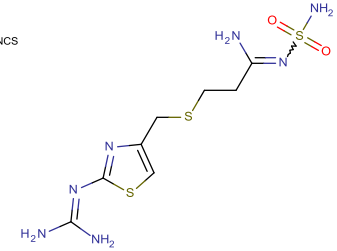
A17



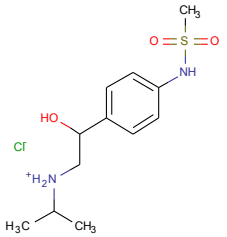
A18



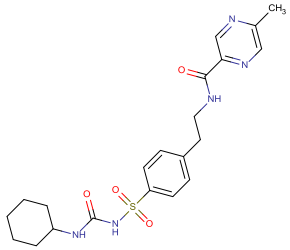
A19



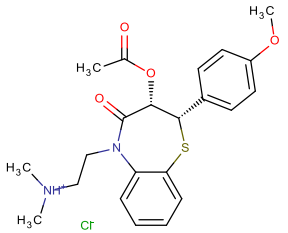
A20



A21

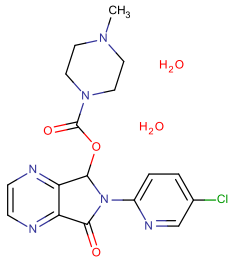


A22

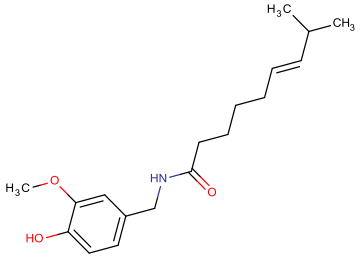


A23

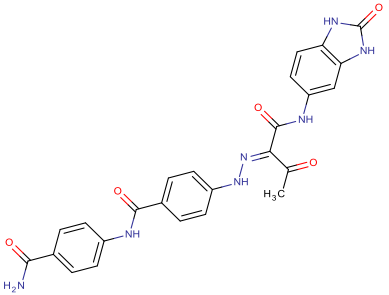
Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1



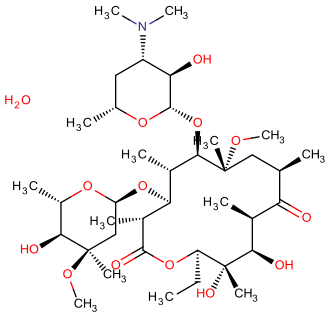
A24



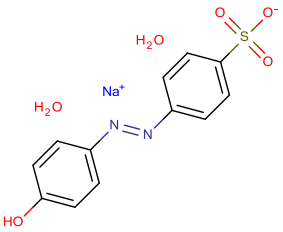
A25



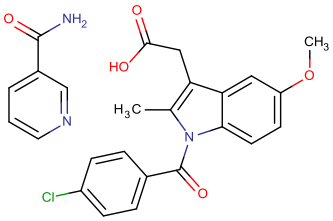
A26



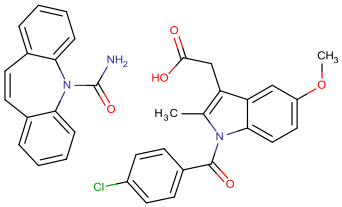
A27



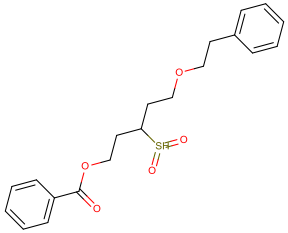
A28



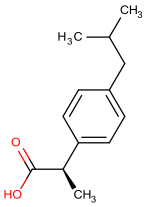
A29



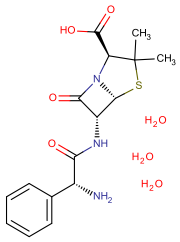
A30



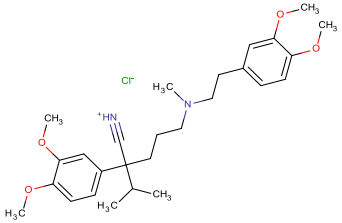
A31



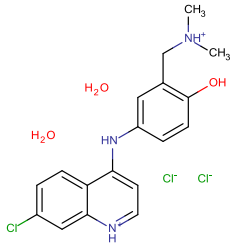
A32



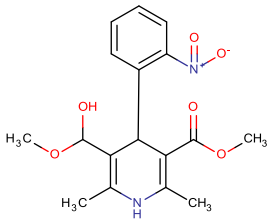
A33



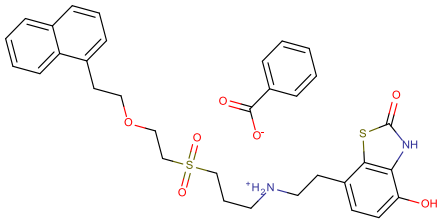
A34



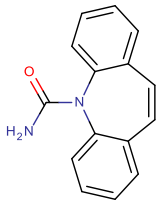
A35



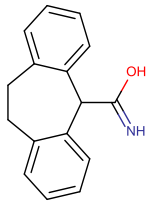
A36



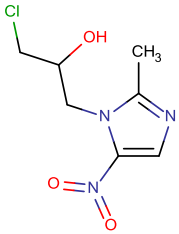
A37



A38

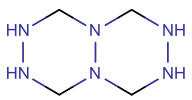


A39

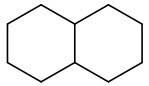


A40

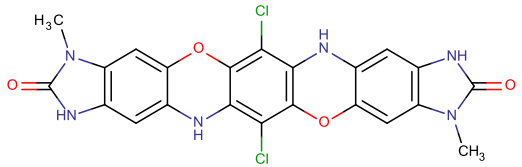
Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1 (continued)



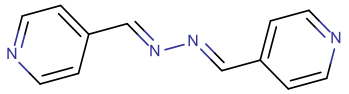
B1



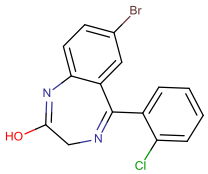
B2



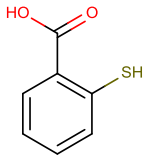
B3



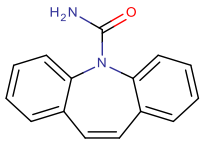
B4



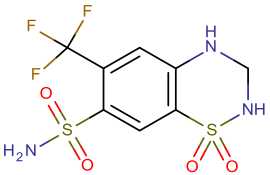
B5



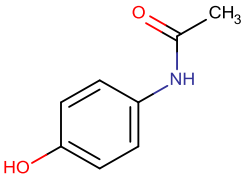
B6



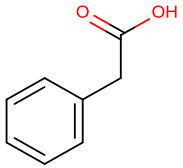
B7



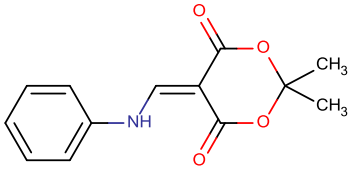
B8



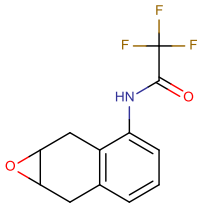
B9/B10



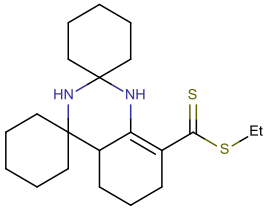
B11



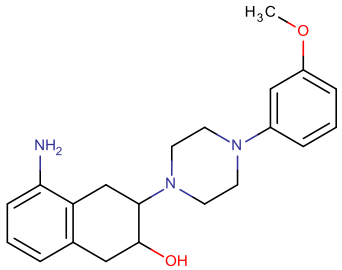
B12



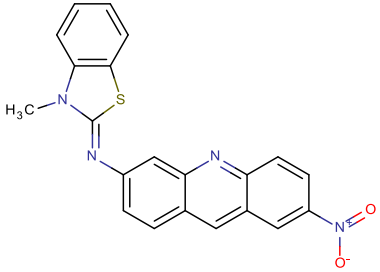
B13



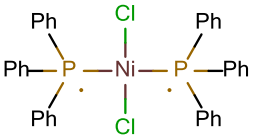
B14



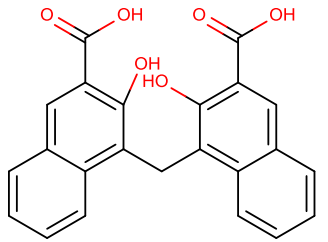
B15



B16

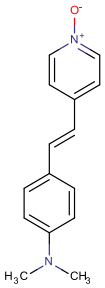


B17

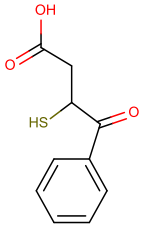


B18

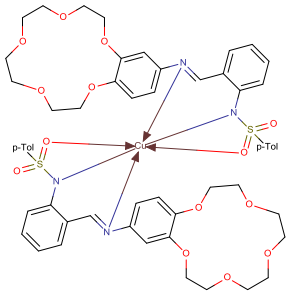
Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1 (continued)



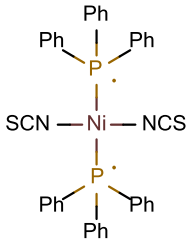
B19



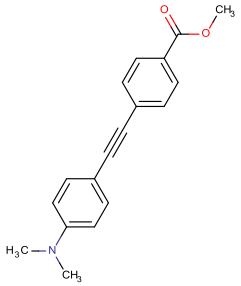
B20



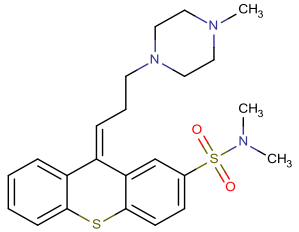
B21



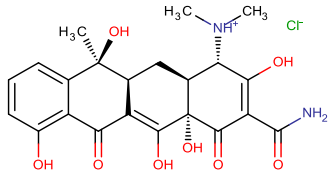
B22



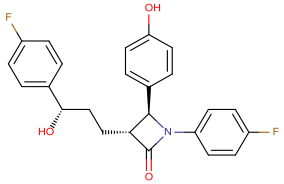
B23



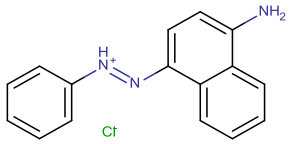
B24



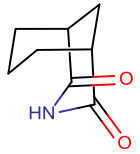
B25



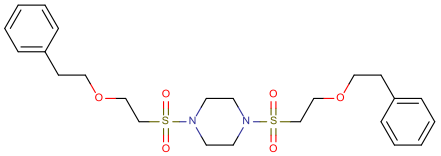
B26



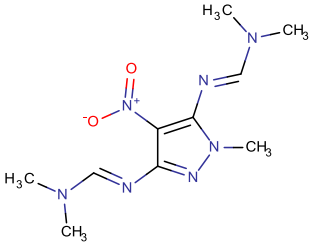
B27



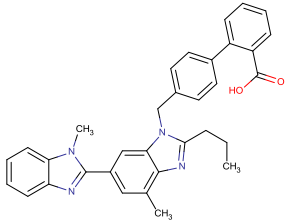
B28



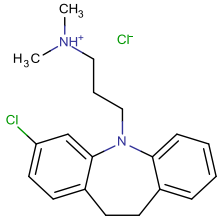
B29



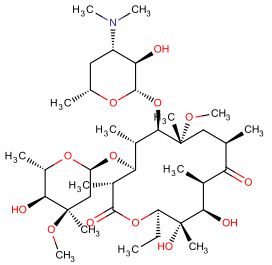
B30



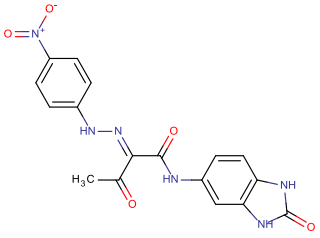
B31/32



B33

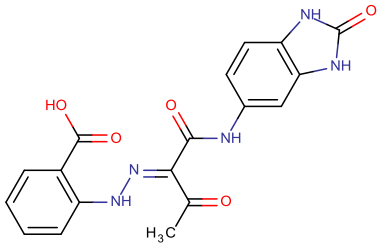


B34

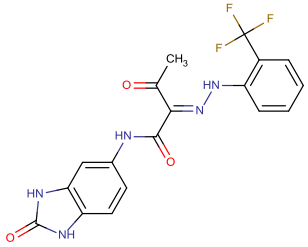


B35

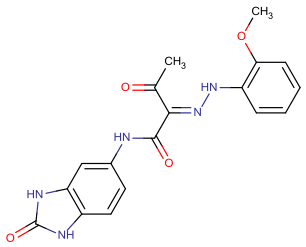
Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1 (continued)



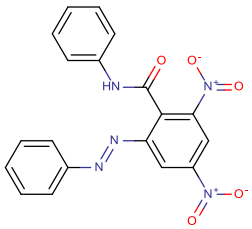
B36



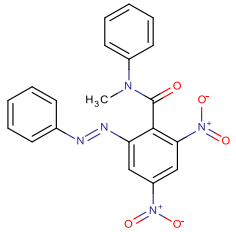
B37



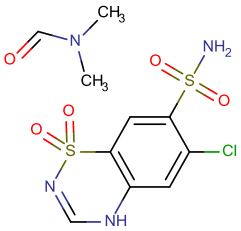
B38



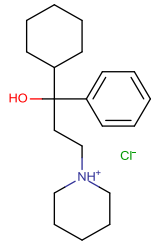
B39



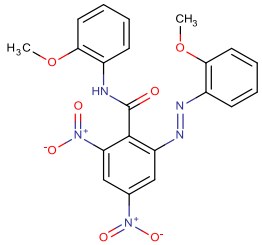
B40



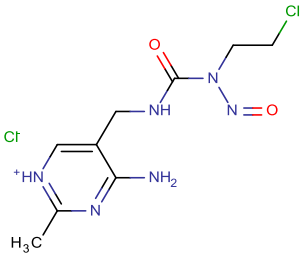
B41



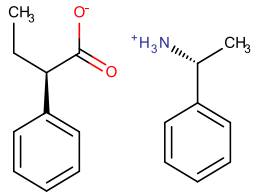
B42



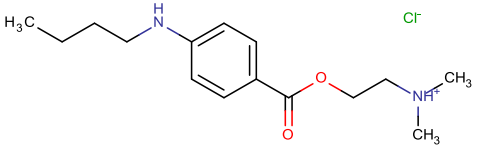
B43



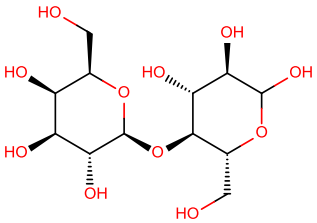
B44



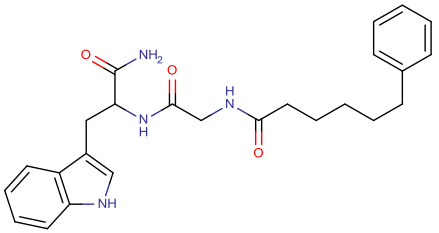
B45/46



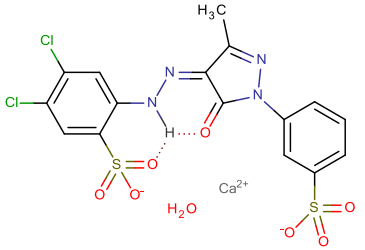
B47



B48

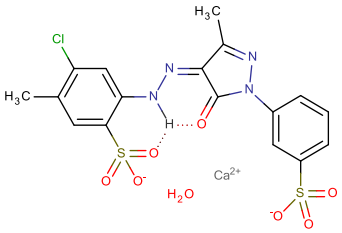


B49

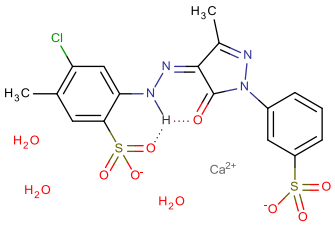


B50

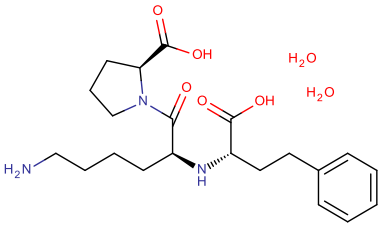
Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1 (continued)



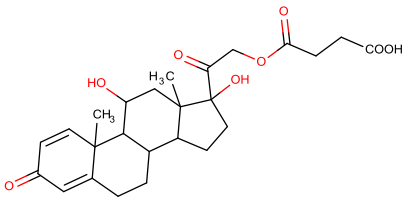
B51



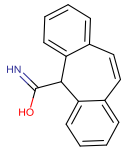
B52



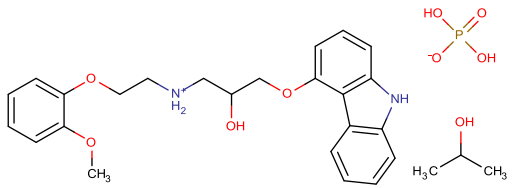
B53



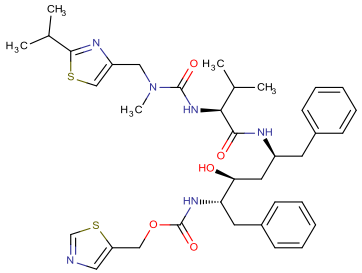
B54



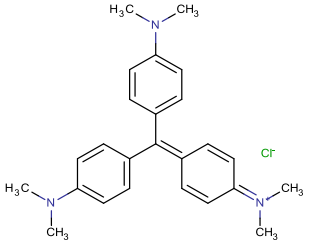
B55



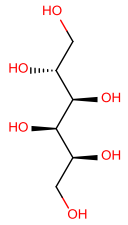
B56



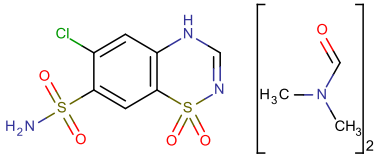
B57



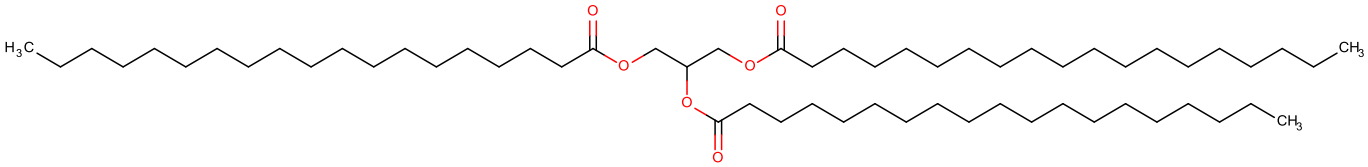
B58



B59



B60



B61

Figure 2.1 Molecular structures of the 101 compounds listed in Table 2.1

2.2 Methods

2.2.1 Software

The crystallographic software used throughout this work is summarised in Table 2.3.

The Z-matrices used in the crystal structure solutions were automatically generated by DASH from previously deposited crystal structure coordinates in MOL2 or CIF format. Unless otherwise stated, all torsion angles, including double and triple bonds[3], were allowed to rotate freely. This was implemented to ensure consistency in the molecular description as well as to maximise the structural complexity of all crystal structures. Water molecules were modeled as a single oxygen atom and hence contribute as only 3 positional DoF. Similarly, hydrochlorides were represented by a single chlorine atom.

The SA parameter optimisation was carried out with the computer program 'irace' 2.14.0 (López-Ibáñez *et al.*, 2011), which is freely downloadable from http://iridia.ulb.ac.be/irace/. 'irace' is implemented as an R-package (Team, 2011), which was obtained from CRAN (Hornik, 2015)

The default settings of the conformer generator were used as supplied by CCDC, to ensure results were consistent with those obtainable by a general user. By default, an ensemble of up to 200 conformers was generated and ranked in order of likelihood as predicted by the program.

Statistical analysis was carried out with Minitab 17.0 (Minitab, 2010).

2.3 Hardware

The work carried out during this project, including all DASH and 'irace' calculations, was performed on the computers that are summarised in Table 2.4.

[3] The Z-matrix generator in DASH relies upon connectivity information present in the input structure in order to make decisions about which non-ring torsion angles in the molecule(s) are free to rotate. As CIFs do not have such connectivity information, then double and triple bonds are erroneously included in the list of free torsions. In general, this leads to only a small increase in the number of parameters to be optimised by DASH and so no effort has been made to eliminate them. Furthermore, in the case of double bonds, it obviates the need to check both *cis* and *trans* configurations.

Table 2.3 Summary of used crystallographic software

Software	Version	Application	Reference
DASH	3.3.2 3.2	Indexing* Space group determination† Pawley refinement Structure solution	(David *et al.*, 2006)
dash.x	3.3.2 3.3.1	Structure solution (under Linux) Irace calculations (under Linux)	CCDC private communications
MDASH	3.1	Structure solution	(Griffin *et al.*, 2009)
Conformer Generator	0.9.3	Generation of likely conformers	CCDC private communications
TOPAS	4.2	Indexing Pawley refinement Rietveld refinement	(Coelho, 2003)
CSD**	5.36	Model building	(Allen, 2002)
MarvinSketch	6.0.5	Model building	(ChemAxon, 2011)
ConQuest	1.17	Structure mining of CSD	(Bruno *et al.*, 2002)
Mercury	3.3	Structure visualisation	(Macrae *et al.*, 2008)
Mogul	1.6	Structure verification	(Bruno *et al.*, 2004)
enCIFer	1.51	CIF verification	(Allen *et al.*, 2004)
PLATON	1.51	Unit cell conversion	(Spek, 2003)

* Via interface to DICVOL91 (Boultif and Louer, 1991); † with ExtSym as implemented in DASH; ** This work considers the CSD a comprehensive database of all published molecular crystal structures. Structures in the Protein Databank (PDB) and Inorganic Crystal Structure Database (ICSD) were outside the scope of this work.

Table 2.4 Hardware summary

Computer Name	CPU	RAM	Operating system
PC1	Intel Core 2 Quad Q9400 (2.66GHz)	4 GB	Windows Enterprise 7 (64-bit)
SR8	2 × Intel Xeon E5520 (2.270GHz)	32 GB	Windows Server 2008 R2 Datacenter (64-bit)
NS	2 × Intel Xeon E5-2630 v2 (2.60GHz)	16 GB	Windows Professional 7 (64-bit)
NS2	2 × Intel Xeon E5-2630 0 (2.30GHz)	16 GB	Windows 7 Enterprise (64-bit) Ubuntu 13.04 (32-bit)

2.4 References

Albov DV, Jassem A and Kuznetsov AI (2006) An independent refinement of H-atom coordinates from laboratory X-ray powder data in tetraformaltrisazine. *Acta Cryst. Sect. E* **62**:o1449-o1451

Alleaume M (1967) PhD Thesis, Bordeaux, France

Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst. Sect. B* **58**:380-388

Allen FH, Johnson O, Shields GP, Smith BR and Towler M (2004) CIF applications. XV. enCIFer: a program for viewing, editing and visualizing CIFs. *J. Appl. Cryst.* **37**:335-338

Assaad T and Rukiah M (2011) Powder X-ray study of racemic (2RS,3RS)-5-amino-3-[4-(3-methoxyphenyl)piperazin-1-yl]-1,2,3,4-tetrahydronaphthalen-2-ol. *Acta Cryst. Sect. C* **67**:o469-o472

Avila EE, Mora AJ, Delgado GE, Contreras RR, Rincon L, Fitch AN and Brunelli M (2009) Structure and conformational analysis of a bidentate pro-ligand, $C_{21}H_{34}N_2S_2$, from powder synchrotron diffraction data and solid-state DFTB calculations. *Acta Cryst. Sect. B* **65**:639-646

Bamgboye TT and Sowerby DB (1986) Crystal structure of $Ni(NCS)_2 \cdot 2Ph_3P$. *Polyhedron* **5**:1487-1488

Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W and Morris J (2001) Ritonavir: An extraordinary example of conformational polymorphism. *Pharm. Res.* **18**:859-866

Beale JP and Stephens NC (1972) X-ray analysis of Th 1165a and salbutamol. *J. Pharm. Pharmacol.* **24**:277

Beko SL, Urmann D, Lakatos A, Glaubitz C and Schmidt MU (2012) Nimustine hydrochloride: the first crystal structure determination of a 2-chloroethyl-N-nitrosourea hydrochloride derivative by X-ray powder diffraction and solid-state NMR. *Acta Cryst. Sect. C* **68**:o144-o148

Borea PA, Gilli G, Bertolasi V and Ferretti V (1987) Stereochemical features controlling binding and intrinsic activity properties of benzodiazepine-receptor ligands. *Mol. Pharmacol.* **31**:334-344

Bortolotti M, Lonardelli I and Pepponi G (2011) Determination of the crystal structure of nifedipine form C by synchrotron powder diffraction. *Acta Cryst. Sect. B* **67**:357-364

Boultif A and Louer D (1991) Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method. *J. Appl. Cryst.* **24**:987-993

Brammer L and Stevens ED (1989) Structure of dichlorobis(triphenylphosphine)nickel(II). *Acta Cryst. Sect. C* **45**:400-403

Bruning J, Alig E and Schmidt MU (2010) Ezetimibe anhydrate, determined from laboratory powder diffraction data. *Acta Cryst. Sect. C* **66**:o341-o344

Bruno IJ, Cole JC, Edgington PR, Kessler M, Macrae CF, McCabe P, Pearson J and Taylor R (2002) New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Cryst. Sect. B* **58**:389-397

Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE and Orpen AG (2004) Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **44**:2133-2144

Burley JC (2005) Structure and intermolecular interactions of glipizide from laboratory X-ray powder diffraction. *Acta Cryst. Sect. B* **61**:710-716

Burley JC, van de Streek J and Stephens PW (2006) Ampicillin trihydrate from synchrotron powder diffraction data. *Acta Cryst. Sect. E* **62**:O797-O799

Bushmarinov IS, Dmitrienko AO, Korlyukov AA and Antipin MY (2012) Rietveld refinement and structure verification using `Morse' restraints. *J. Appl. Cryst.* **45**:1187-1197

Carpy A, Leger JM and Melchiorre C (1985) Structure of α-isopropyl-α-(N-methyl-n-homoveratryl)-γ-aminopropyl-3,4- dimethoxyphenylacetonitrile hydrochloride, verapamil, $C_{27}H_{38}N_2O_4$.HCl. *Acta Cryst. Sect. C* **41**:624-627

ChemAxon (2011) Marvin 5.4.1.1, http:www.chemaxon.com, Oct 2015

Chernyshev VV, Kukushkin SY and Velikodny YA (2010) Carvedilol dihydrogen phosphate propan-2-ol solvate from powder diffraction data. *Acta Cryst. Sect. E* **66**:o613

Chernyshev VV, Tafeenko VA, Makarov VA, Sonneveld EJ and Schenk H (2000) 3,5-Bis[(N,N-dimethylamino)methyleneamino]-1-methyl-4-nitropyrazole from X-ray powder diffraction data. *Acta Cryst. Sect. C* **56**:1159-1160

Chernyshev VV, Yatsenko AV, Kuvshinov AM and Shevelev SA (2002) Unexpected molecular structure from laboratory powder diffraction data. *J. Appl. Cryst.* **35**:669-673

Clegg W and Teat SJ (2000) Tetracycline hydrochloride: a synchrotron microcrystal study. *Acta Cryst. Sect. C* **56**:1343-1345

Coelho A (2003) TOPAS user manual. Bruker AXS GmbH, Karlsruhe,

David WIF, Shankland K and Shankland N (1998) Routine determination of molecular crystal structures from powder diffraction data. *Chem. Comm.*:931-932

David WIF, Shankland K, van de Streek J, Pidcock E, Motherwell WDS and Cole JC (2006) DASH: a program for crystal structure determination from powder diffraction data. *J. Appl. Cryst.* **39**:910-915

Dinnebier RE, Sieger P, Nar H, Shankland K and David WIF (2000) Structural characterization of three crystalline modifications of telmisartan by single crystal and high-resolution X-ray powder diffraction. *J. Pharm. Sci.* **89**:1465-1479

Donaldson JD, Leary JR, Ross SD, Thomas MJK and Smith CH (1981) The structure of the orthorhombic form of tolbutamide (1-N-butyl-3-P-toluenesulphonylurea). *Acta Cryst. Sect. B* **37**:2245-2248

Dorokhov AV, Chernyshov DY, Burlov AS, Garnovskii AD, Ivanova IS, Pyatova EN, Tsivadze AY, Aslanov LA and Chernyshev VV (2007) Synchrotron powder diffraction in a systematic study of 4'- 2-(tosylamino)benzylideneamino -2,3benzo-15-crown-5 complexes. *Acta Cryst. Sect. B* **63**:402-410

Dupont L and Dideberg O (1972) Crystal-structure of hydrochlorothiazide, $C_7H_8ClN_3O_4S_2$. *Acta Cryst. Sect. B* **B 28**:2340

Eibl S, Fitch AN, Brunelli M, Evans AD, Pattison P, Plazanet M, Johnson MR, Alba-Simionesco C and Schober H (2009) trans-Decahydronaphthalene (decalin) from powder diffraction data. *Acta Cryst. Sect. C* **65**:o278-o280

Fernandes P, Florence A, Shankland K, Karamertzanis PG, Hulme AT and Anandamanoharan P (2007a) Powder study of (R)-1-phenylethylammonium (R)-2-phenylbutyrate form 2. *Acta Cryst. Sect. E* **63**:o247-o249

Fernandes P, Florence AJ, Shankland K, Karamertzanis PG, Hulme AT and Anandamanoharan RP (2007b) Powder study of (R)-1-phenylethylammonium (R)-2-phenylbutyrate form 3. *Acta Cryst. Sect. E* **63**:o202-o204

Fernandes P, Florence AJ, Shankland K, Shankland N and Johnston A (2006) Powder study of chlorothiazide N,N-dimethylformamide solvate. *Acta Cryst. Sect. E* **62**:o2216-o2218

Fernandes P, Shankland K, Florence AJ, Shankland N and Johnston A (2007c) Solving molecular crystal structures from X-ray powder diffraction data: The challenges posed by γ-carbamazepine and chlorothiazide N,N,-dimethylformamide (1/2) solvate. *J. Pharm. Sci.* **96**:1192-1202

Florence AJ, Baumgartner B, Weston C, Shankland N, Kennedy AR, Shankland K and David WIF (2003) Indexing powder patterns in physical form screening: Instrumentation and data quality. *J. Pharm. Sci.* **92**:1930-1938

Florence AJ, Shankland K, Gelbrich T, Hursthouse MB, Shankland N, Johnston A, Fernandes P and Leech CK (2008) A catemer-to-dimer structural transformation in cyheptamide. *CrystEngComm* **10**:26-28

Florence AJ, Shankland N, Shankland K, David WIF, Pidcock E, Xu XL, Johnston A, Kennedy AR, Cox PJ, Evans JSO, Steele G, Cosgrove SD and Frampton CS (2005) Solving molecular crystal structures from laboratory X-ray powder diffraction data with DASH: the state of the art and challenges. *J. Appl. Cryst.* **38**:249-259

Freer AA, Bunyan JM, Shankland N and Sheen DB (1993) Structure of (S)-(+)-ibuprofen. *Acta Cryst. Sect. C* **49**:1378-1380

Fries DC, Rao ST and Sundaral M (1971) Structural chemistry of carbohydrates.3. Crystal and molecular structure of 4-o-β-d-galactopyranosyl-α-d-glucopyranose monohydrate α-lactose monohydrate. *Acta Cryst. Sect. B* **B 27**:994

Fujinaga M and James MNG (1980) SQ-14,225 - 1-(D-3-mercapto-2-methylpropionyl)-L-proline. *Acta Cryst. Sect. B* **36**:3196-3199

Gadret M, Goursolle M, Leger JM, Colleter JC and Carpy A (1976) Crystal-structure of sotalol hydrochloride - para-(1-hydroxy-2-isopropylaminoethyl)methanesulfonanilide. *Acta Cryst. Sect. B* **32**:2757-2761

Griffin TAN, Shankland K, van de Streek J and Cole J (2009) MDASH: a multi-core-enabled program for structure solution from powder diffraction data. *J. Appl. Cryst.* **42**:360-361

Haynes DA, van de Streek J, Burley JC, Jones W and Motherwell WDS (2006) Pamoic acid determined from powder diffraction data. *Acta Cryst. Sect. E* **62**:o1170-o1172

Helmholdt RB, Peschar R and Schenk H (2002) Structure of C15-, C17- and C19-mono-acid β-triacylglycerols. *Acta Cryst. Sect. B* **58**:134-139

Himes VL, Mighell AD and De Camp WH (1981) Structure of carbamazepine: 5H-dibenz[b,f]azepine-5-carboxamide. *Acta Cryst. Sect. B* **37**:2242-2245

Hodgson DJ and Asplund RO (1991) Phenylacetic acid. *Acta Cryst. Sect. C* **47**:1986-1987

Hornik K (2015) The R FAQ, https://CRAN.R-project.org/doc/FAQ/R-FAQ.html, Oct 2015

Hulme AT, Fernandes P, Florence A, Johnston A and Shankland K (2006) Powder study of 3-azabicyclo[3.3.1]nonane-2,4-dione form 2. *Acta Cryst. Sect. E* **62**:o3046-o3048

Ivashevskaja SN, Aleshina LA, Andreev VP, Nizhnik YP, Chernyshev VV and Schenk H (2003) 4-(4'-Dimethylaminostyryl)pyridine N-oxide from powder data. *Acta Cryst. Sect. E* **59**:o1006-o1008

Ivashevskaya SN, van de Streek J, Djanhan JE, Bruening J, Alig E, Bolte M, Schmidt MU, Blaschka P, Hoeffken HW and Erk P (2009) Structure determination of seven phases and solvates of Pigment Yellow 183 and Pigment Yellow 191 from X-ray powder and single-crystal data. *Acta Cryst. Sect. B* **65**:212-222

Johnston A, Florence AJ, Shankland K, Markvardsen A, Shankland N, Steele G and Cosgrove SD (2004) Powder study of N- 2-(4-hydroxy-2-oxo-2,3-dihydro-1,3-benzothiazol-7-yl)ethyl -3- 2-(2-n aphthalen-1-ylethoxy)ethylsulfonyl propylaminium benzoate. *Acta Cryst. Sect. E* **60**:O1751-O1753

Kato Y, Haimoto Y and Sakurai K (1979) Refinement of the crystal-structure of creatine monohydrate. *Bull. Chem. Soc. Jpn.* **52**:233-234

Kennedy AR, Hughes MP, Monaghan ML, Staunton E, Teat SJ and Smith WE (2001) Supramolecular motifs in s-block metal bound sulfonated monoazo dyes. *Dalton Trans.*:2199-2205

Kojicprodic B, Ruzictoros Z, Sunjic V, Decorte E and Moimas F (1984) Absolute conformation and configuration of (2S, 3S)-3-acetoxy-5-(dimethylaminoethyl)-2-(4-methoxyphenyl)-2,3-dihydro-1,5 -benzothiazepin-4(5H)-one chloride (dilthiazem hydrochloride). *Helv. Chim. Acta* **67**:916-926

Koo CH, Cho SI and Yeon YH (1980) Crystal and molecular structure of chlorpropamide. *Arch. Pharmacal Res.* **3**:37-50

Lefebvre J, Willart J-F, Caron V, Lefort R, Affouard F and Danede F (2005) Structure determination of the 1/1 α/β mixed lactose by X-ray powder diffraction. *Acta Cryst. Sect. B* **61**:455-463

Llinas A, Fabian L, Burley JC, van de Streek J and Goodman JM (2006) Amodiaquinium dichloride dihydrate from laboratory powder diffraction data. *Acta Cryst. Sect. E* **62**:o4196-o4199

López-Ibáñez M, Dubois-Lacoste J, Stützle T and Birattari M (2011) The irace package, Iterated Race for Automatic Algorithm Configuration. Université libre de Bruxelles, IRIDIA, TR/IRIDIA/2011-004, Belgium, http://iridia.ulb.ac.be/irace/

Maccaroni E, Malpezzi L and Masciocchi N (2010) Trihexyphenidyl hydrochloride: a powder diffraction study. *Acta Cryst. Sect. E* **66**:o2511

Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J and Wood PA (2008) Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. *J. Appl. Cryst.* **41**:466-470

Majumder M, Buckton G, Rawlinson-Malone CF, Williams AC, Spillman MJ, Pidcock E and Shankland K (2013) Application of hydrogen-bond propensity calculations to an indomethacin-nicotinamide (1:1) co-crystal. *CrystEngComm* **15**:4041-4044

Marder TC (2004) Unpublished single-crystal data.

Minitab (2010) Minitab 17 Statistical Software, Minitab, Inc., State College, PA. www.minitab.com, Oct 2015

Nichols G and Frampton CS (1998) Physicochemical characterization of the orthorhombic polymorph of paracetamol crystallized from solution. *J. Pharm. Sci.* **87**:684-693

Nishibori E, Ogura T, Aoyagi S and Sakata M (2008) Ab initio structure determination of a pharmaceutical compound, prednisolone succinate, from synchrotron powder data by combination of a genetic algorithm and the maximum entropy method. *J. Appl. Cryst.* **41**:292-301

Noguchi S, Fujiki S, Iwao Y, Miura K and Itai S (2012a) Clarithromycin monohydrate: a synchrotron X-ray powder study. *Acta Cryst. Sect. E* **68**:o667-668

Noguchi S, Miura K, Fujiki S, Iwao Y and Itai S (2012b) Clarithromycin form I determined by synchrotron X-ray powder diffraction. *Acta Cryst. Sect. C* **68**:O41-O44

Nowell H, Attfield JP, Cole JC, Cox PJ, Shankland K, Maginn SJ and Motherwell WDS (2002) Structure solution and refinement of tetracaine hydrochloride from X-ray powder diffraction data. *New J. Chem.* **26**:469-472

Post ML and Horn AS (1977) The crystal and molecular structure of the tricyclic antidepressant chlorimipramine hydrochloride: 3-chloro-5-(3-dimethylaminopropyl)-10,11-dihydro-5H-dibenz[b,f]azepine hydrochloride. *Acta Cryst. Sect. B* **33**:2590-2595

Rohlicek J, Maixner J, Pazout R, Husak M, Cibulkova J and Kratochvil B (2010) Alaptide from synchrotron powder diffraction data. *Acta Cryst. Sect. E* **66**:O821-U2229

Rukiah M and Al-Ktaifani M (2011) 2-(Benzoylsulfanyl)acetic acid and 2,5-dioxopyrrolidin-1-yl 2-(benzoylsulfanyl)acetate by powder X-ray diffraction studies. *Acta Cryst. Sect. C* **67**:o166-o170

Rukiah M and Assaad T (2010) 2,2,2-Trifluoro-N-(1a,2,7,7a-tetrahydronaphtho[2,3-b]oxiren-3-yl)acetamide by X-ray powder diffraction. *Acta Cryst. Sect. C* **66**:o475-o478

Rukiah M, Lefebvre J, Hernandez O, van Beek W and Serpelloni M (2004) Ab initio structure determination of the γ- form of d-sorbitol (d-glucitol) by powder synchrotron X-ray diffraction. *J. Appl. Cryst.* **37**:766-772

Schmidt MU, Ermrich M and Dinnebier RE (2005) Determination of the structure of the violet pigment $C_{22}H_{12}C_{l2}N_6O_4$ from a non-indexed X-ray powder diagram. *Acta Cryst. Sect. B* **61**:37-45

Sergeev GB, Sergeev BM, Morosov YN and Chernyshev VV (2010) β-Polymorph of phenazepam: a powder study. *Acta Cryst. Sect. E* **66**:o2623

Shankland N, David WIF, Shankland K, Kennedy AR, Frampton CS and Florence AJ (2001) Structural transformations in zopiclone. *Chem. Comm.* 10.1039/b107075d:2204-2205

Shankland N, Love SW, Watson DG, Knight KS, Shankland K and David WIF (1996) Constrained Rietveld refinement of β-H-1(1) decadeuteriodopamine deuteriobromide using powder neutron diffraction data. *Faraday Trans.* **92**:4555-4559

Shanmuga Sundara Raj S, Fun H-K, Zhang J, Xiong R-G and You X-Z (2000) Pyridine-4-carbaldehyde azine. *Acta Cryst. Sect. C* **56**:e274-e275

Shin HS, Song H, Kim E and Chung KB (1995) The crystal and molecular-structure of 1-(3-chloro-2-hydroxypropyl)-2-methyl-5-nitroimidazole (ornidazole), $C_7H_{10}ClN_3O_3$. *Bull. Korean Chem. Soc.* **16**:912-915

Smrcok L, Jorik V, Scholtzova E and Milata V (2007) Ab initio structure determination of 5-anilinomethylene-2,2-dimethyl-1,3-dioxane-4,6-dione from laboratory powder data - a combined use of X-ray, molecular and solid-state DFT study. *Acta Cryst. Sect. B* **63**:477-484

Sorrenti M, Catenacci L, Cruickshank DL and Caira MR (2013) Lisinopril dihydrate: Single-crystal X-ray structure and physicochemical characterization of derived solid forms. *J. Pharm. Sci.* **102**:3596-3603

Spek A (2003) Single-crystal structure validation with the program PLATON. *J. Appl. Cryst.* **36**:7-13

Steiner T (2000) S-H...S hydrogen-bond chain in thiosalicylic acid. *Acta Cryst. Sect. C* **56**:876-877

Team RDC (2011) R: a language and environment for statistical computing, R Foundation for Statistical Computing, http://www.R-project.org/

Vallcorba O, Latorre S, Alcobe X, Miravitlles C and Rius J (2011) (Z)-3-Methyl-N-(7-nitroacridin-3-yl)-2,3-dihydro-1,3-benzothiazol-2-imine from laboratory powder diffraction data. *Acta Cryst. Sect. C* **67**:o425-o427

van de Streek J, Bruening J, Ivashevskaya SN, Ermrich M, Paulus EF, Bolte M and Schmidt MU (2009) Structures of six industrial benzimidazolone pigments from laboratory powder diffraction data. *Acta Cryst. Sect. B* **65**:200-211

Yatsenko AV, Chernyshev VV, Paseshnichenko KA and Schenk H (2001) 4-(Phenyldiazenyl)naphthalen-1-amine and its hydrochloride. *Acta Cryst. Sect. C* **57**:295-297

3 The 101 data sets: selection criteria and baseline DASH performance

3.1 Introduction

A study carried out by Florence *et al*. (2005) on 35 industrially relevant compounds concluded that crystal structures with greater than 20 DoF can be classed as 'complex' and were representative of the current limits of SDPD at the time (Figure 3.1). In contrast, crystal structures with up to 15 DoF solved easily, with good accuracy and reproducibility. For the examples with structural complexity of 15-20 DoF, a reduced but still reasonable success rate was characteristic. This conclusion remains in good agreement with the observations based on the CSD entries (Figures 1.7) and laid the foundation of the work carried out in this thesis. With an overall aim of increasing the complexity limits of SDPD, the crucial first step was to assemble a data set to serve three main purposes:

1) to better populate the mid-range of structural complexity (15 ≤ DoF ≤ 20) shown in Figure 3.1, and to introduce crystal structures of up to 30 DoF;
2) to establish baseline DASH performance against a wide variety of structures;
3) to evaluate the efficiency of the enhanced approaches to SDPD outlined in section 1.7;

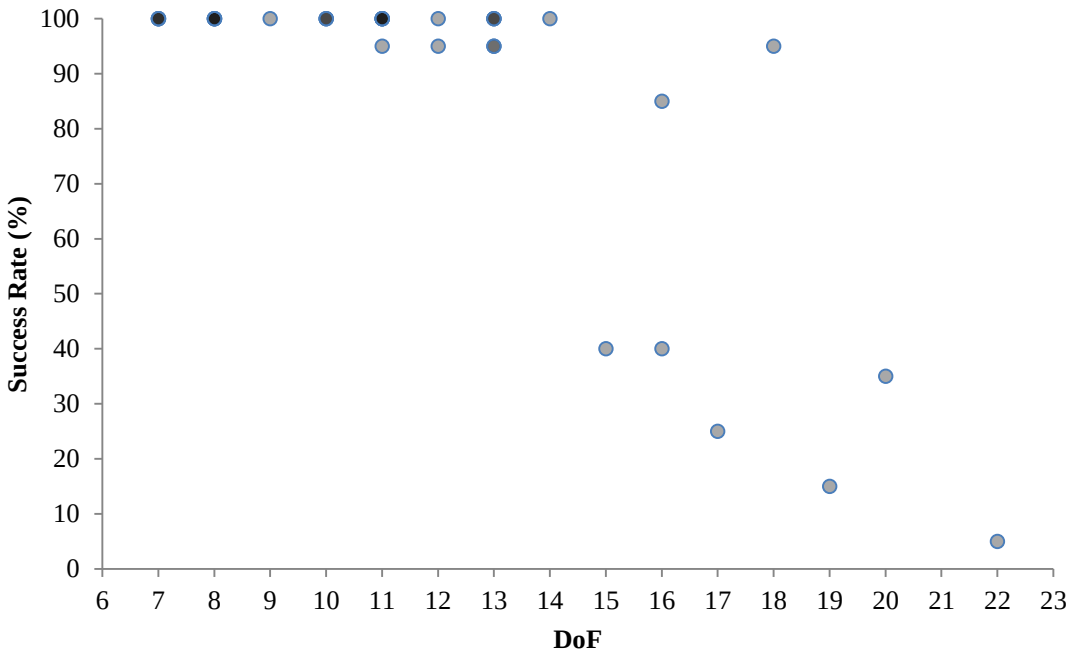


Figure 3.1 A graph of the success rate (SR) vs DoF, based on the results published by Florence *et al.* (2005). The depth of shading of each point is proportional to the number of structures representative of a particular DoF. The simplicity of the crystal structures with less than 15 DoF is demonstrated by the high SR (approximately 100%), whilst the region between 15 and 20 DoF is characterised by reduced, but still acceptable, SRs. The one example with DoF greater than 20 shows significant reduction in its SR.

A data set size of 100 molecules was considered sufficiently large to facilitate a comprehensive and systematic study. Such a large sample size accommodated multiple representatives of each of the possible DoF, allowing to study the effect which single- and multi- component representatives of particular DoFs have on the crystal structure solution.

Since the data for the 35 crystal structures used by Florence *et al*. were freely available from http://www.powderdata.net, they formed the base for the final data set (FDS). The remaining 65 structures were assembled using in-house data and data associated with crystal structures published in International Union of Crystallography (IUCr) journals (see Table 2.1). IUCr publications were specifically targeted as being the most likely to include high-quality diffraction data / metadata[4] along with published crystal structures.

The selection criteria for the dataset assembly were:

1. Small organic molecules were targeted[5], as the class of compounds DASH is best suited to solving.
2. Their crystal structures should previously have been solved from powder X-ray diffraction data. The importance of this was twofold: a) structures would be indicative of crystallographic and industrial interest and hence would accurately illustrate the up-to-date limits of SDPD; and b) the published structures could be used to validate the performance of DASH and to assess the quality of the solutions.
3. The powder X-ray diffraction data should be available. Whilst, in principle, data from interesting compounds could be collected 'in-house' if necessary, this would have increased the work load significantly and there is no guarantee that the crystal form reported would be the crystal form purchased.

   The possibility of using simulated powder diffraction data was briefly considered, but rejected as being open to criticism on the basis that, no matter how carefully it is synthesised, it is not truly representative of real-world data.

It is worth noting that whilst representatives with greater than 15 DoF were of particular interest (from the standpoint of expanding the complexity of crystal structures which can be solved from PXRD using DASH), the presence of structures of lower complexity was also essential.

[4] Metadata are the useful data terms that are used to describe the raw diffraction data, such as wavelength, instrument geometry, sample preparation details, count times etc,
[5] 'Small molecule' is used here in the commonly-accepted crystallography sense *i.e*. not proteins

They were used to ensure that any changes made to the SA algorithm to improve performance at high complexity, did not compromise efficacy at relatively low complexity. Additionally, the less complex examples were required for the parameter tuning calculations, ensuring a sufficient number of runs were capable of reaching the global minima and hence improving the chances of successful SA parameter optimisation.

In total, diffraction data associated with 101 crystal structures were successfully assembled. For the purpose of the SA parameter optimisation (Chapter 4), these were divided into two subsets – the 'training' and 'test' sets. The training set was a representative sample of 40 structures which, in the parameter tuning experiments, were used to benchmark the performance of DASH and optimise its SA parameters. The remaining 61 structures comprised the test set, which was used to then independently validate these optimised SA parameters.

3.2 Experimental

3.2.1 Data set treatments

A summary of the crystallographic information associated with each selected crystal structure is presented in Table 3.1. Each data set required a number of preparation steps, the first of which was the conversion of all powder diffraction data files into the same ".xye" format. The number of different formats used by authors depositing data with their crystal structures prevented routine file conversion using programmes such as PowDLL (Kourkoumelis, 2013) or Python scripts. Consequently, all files were manually prepared to ensure that no artefacts were inadvertently introduced into the data.

Subsequently, the standard sequence of steps of crystal structure solution was followed (Figure 1.1). First, the unit cell, space group and extracted integrated intensities (Pawley fit) were determined using DASH. Data sets for which these steps were unsuccessful were rejected. Whilst this could be considered to introduce some bias into the eventual data set used for this work, it is worth remembering that the work is focussed on improving the performance of DASH, as a representative of the simulated annealing approach to global optimisation. As such, data which present a challenge during the indexing or Pawley refinement stages, especially due to poor data quality, can be legitimately excluded.

Prior to the crystal structure solution step, a rigid-body Rietveld refinement of the reported crystal structure was performed using DASH in order to establish the 'target' χ^2 value for the

simulated annealing *i.e.* a χ^2 value representative of the solved crystal structure. This value was used as a guide to help with the rapid identification of correct crystal structure solutions identified by DASH, and provided a useful guide for SA parameter optimisation. The key data associated with the 101 crystal structures are to be found in Appendix A, including the .xye data files "as received" and all relevant DASH files for the Pawley and SA steps.

3.2.2 Crystal structure solution

Initially, 50 SA runs were executed on all 101 structures, using the default DASH parameters (T_0 = 0; N_1 = 20; N_2 = 25). Each run was set to perform 10^7 SA steps followed by a short simplex calculation. A χ^2 multiplier of 1 ensured the full number of SA steps was always carried out and the SA was not prematurely terminated. The starting molecular conformers were randomly generated and all of the torsion angles were allowed to rotate freely (from 0^o to 360^o) during the SA calculations. Successful solutions were identified on the basis of their χ^2 value and further confirmed by comparison of coordinates with the reference crystal structure. The four crystal structures for which no reference structures had been previously deposited (A4, A6, B23 and B58), were considered successfully solved when a favourable value of the χ^2 ratio had been achieved[6] (typically in the range from 2 to 10) and the crystal structure was found to make chemical and crystallographic sense.

A March-Dollase correction (in the appropriate direction) was introduced in the SA step for some structures (Table 3.1), in order to take account of intensity distortions due to preferred orientation of the crystallites in the samples.

For crystal structures which were not solved with the initial 50 SA runs, an additional 100 SA runs of 10^7 steps were performed. If a structure remained unsolved after this further set of runs, a final attempt at a crystal structure solution was performed with another 500 SA runs of 5×10^7 steps. In order to speed up these longer calculations, the 500 runs were performed using MDASH to spread the calculations over 10 CPU cores *i.e.* a batch size of 50 SA runs. Those structures which still remained unsolved were considered to have a 0% SR.

For consistency and to facilitate comparison of success rates, the same values (315 and 159) of the random seeds in DASH were used throughout. The seeds for the batch 500 SA runs were automatically generated by DASH, starting from the default seed 315 and 159, and increased

[6] $\chi^2_{Ratio} = \chi^2_{Profile}/\chi^2_{Pawley}$

in increments of 50, *i.e.* the second set of 50 runs had the random seeds of 365 and 209, the third set 415 and 259 and so on.

It is also important to note that, whilst deposited crystal structures were used as the starting point for Z-matrices for the majority of the DASH calculations, starting values of the flexible torsion angles were always randomised by DASH and so no advantage (other than the use of good quality bond lengths and bond angles) was conferred by this approach. Indeed, it represents the recommended approach in SDPD of using the most accurate starting model that is available.

Table 3.1 Crystallographic information of the FDS as previously reported**.** λ = wavelength of radiation used in data collection; PO = preferred orientation direction. A-codes represent the training set and B-codes the test(validation) set.

No	Space Group	a (Å)	b (Å)	c (Å)	α (^{0})	β (^{0})	γ (^{0})	Volume (Å^3)	Z'	Total DoF	DoF Position	DoF Orient	DOF Torsion	λ (Å)	PO
A1	*P* $2_12_12_1$	21.14	7.22	6.15	90	90	90	938.41	1	6	3	3	0	0.79984	
A2	*P* 2_1	10.01	8.51	7.40	90	111.74	90	587.47	1	7	3	3	1	1.54056	
A3	*P* $2_12_12_1$	25.54	8.06	5.76	90	90	90	1190.64	1	8	3	3	2	1.54056	
A4	*P* $2_1/c$	8.88	5.41	19.47	90	101.66	90	916.25	1	9	3	3	3	1.54056	
A5	*P* $2_12_12_1$	8.81	17.98	6.84	90	90	90	1083.37	1	10	3	3	4	1.54056	
A6	*P* 2_1	7.57	5.91	14.15	90	95.33	90	630.45	1	10	3	3	4	1.54056	
A7	*P* $2_12_12_1$	5.57	8.85	35.68	90	90	90	1758.13	1	10	3	3	4	0.8000	
A8	*P* $2_1/a$	7.55	14.42	10.25	90	109.60	90	1051.59	1	11	6	3	2	1.54056	
A9	*P b c a*	21.65	8.80	14.56	90	90	90	2774.81	1	11	3	3	5	1.54056	
A10	*P b c* 2_1	10.67	11.48	7.94	90	90	90	972.28	1	11	6	3	2	1.54056	
A11	*P* $2_12_12_1$	9.07	5.22	26.60	90	90	90	1258.54	1	12	3	3	6	1.54056	
A12	*P* $2_1/c$	12.51	5.05	12.19	90	108.90	90	728.36	1	12	6	3	3	1.54056	
A13	*P* $\bar{1}$	6.52	8.53	12.92	84.33	80.58	69.19	661.22	1	12	3	3	6	1.54056	
A14	*P* 2_1	7.98	21.56	4.82	90	109.57	90	782.29	1	13	6	3	4	1.54056	
A15	*P* $2_1/c$	11.81	11.49	13.43	90	111.72	90	1692.28	1	13	6	3	4	1.54056	
A16	*P n a* 2_1	20.22	7.83	9.09	90	90	90	1439.55	1	13	3	3	7	1.54056	
A17	*C m c a*	19.63	4.84	28.80	90	90	90	2738.11	0.5	13	9	3	2	1.54056	
A18	*P* $\bar{1}$	8.65	9.12	11.38	74.72	81.60	88.98	856.78	1	14	3	3	8	0.5200	
A19	*P* $2_1/c$	9.29	23.01	15.28	90	108.06	90	3106.11	1	14	3	3	8	1.54056	
A20	*P* $2_1/c$	17.65	5.29	18.26	90	123.55	90	1421.84	1	15	3	3	9	1.54056	
A21	*C* 2/*c*	15.35	13.48	15.30	90	91.45	90	3164.83	1	15	6	3	6	0.85075	
A22	*P* $\bar{1}$	9.15	24.29	5.18	93.12	101.15	83.48	1121.18	1	16	3	3	10	1.7900	
A23	*P* $2_12_12_1$	12.83	13.06	13.83	90	102.68	90	2262.19	1	16	6	3	7	1.54056	
A24	*P* $2_1/c$	16.37	7.03	17.18	90	108.62	90	1874.61	1	16	9	3	4	1.54056	
A25	*P* $2_1/c$	12.22	14.79	9.47	90	93.98	90	1707.74	1	17	3	3	11	1.54056	

No	Space Group	a (Å)	b (Å)	c (Å)	α (^{0})	β (^{0})	γ (^{0})	Volume (Å^3)	Z'	Total DoF	DoF Position	DoF Orient	DOF Torsion	λ (Å)	PO
A26	*P* $2_1/c$	22.55	4.96	21.28	90	109.45	90	2246.15	1	17	3	3	11	1.54056	
A27	*P* $2_12_12_1$	15.7	18.88	15.03	90	90	90	4454.53	1	17	6	3	8	1.30000	
A28	*P b c n*	14.38	5.81	32.89	90	90	90	2750.03	1	18	12	3	3	1.54056	
A29	*P* $2_1/c$	17.20	5.02	27.38	90	97.31	90	2342.68	1	18	6	6	6	1.54056	
A30	*P* $2_1/c$	10.24	29.15	10.21	90	106.64	90	2921.62	1	18	6	6	6	1.54056	
A31	*P* $2_1/n$	5.07	37.85	9.64	90	97.86	90	1833.22	1	18	3	3	12	1.54056	
A32	*P* 2_1	12.46	8.03	13.54	90	112.89	90	1248.93	2	20	6	6	8	1.54056	
A33	*P* $2_12_12_1$	15.52	18.93	6.67	90	90	90	1960.60	1	20	12	3	5	0.70030	
A34	*P* $\bar{1}$	7.09	10.59	19.20	100.10	93.73	101.55	1382.06	1	22	6	3	13	1.54056	
A35	*P* $2_1/c$	7.84	26.99	10.81	90	92.96	90	2283.7	1	24	15	3	6	1.79000	
A36	*P* $\bar{1}$	9.864	13.89	14.29	61.23	79.83	81.78	1685.37	2	24	6	6	12	0.50000	
A37	*P* $\bar{1}$	7.63	13.67	15.81	84.39	87.47	75.71	1589.52	1	25	6	6	13	1.54056	
A38	*P* $\bar{1}$	5.186	20.58	22.24	84.19	87.98	85.11	2351.44	4	28	12	12	4	0.51561	
A39	*P* $\bar{1}$	5.649	19.56	22.07	84.22	88.41	83.60	2411.72	4	28	12	12	4	1.54056	
A40	*P* $\bar{1}$	13.60	14.05	8.913	71.59	78.73	64.86	1460.09	3	30	9	9	12	0.65278	
B1	*P* $2_1/n$	6.32	4.86	11.33	90	92.04	90	348.32	0.5	6	3	3	0	1.54056	
B2	*P* $2_1/n$	7.81	10.47	5.26	90	90.99	90	430.32	0.5	6	3	3	0	0.69400	
B3	*P* $\bar{1}$	4.28	8.31	14.09	107.23	93.53	97.17	471.94	0.5	6	3	3	0	1.54056	
B4	*P* $2_1/c$	3.85	11.02	12.73	90	92.31	90	539.88	0.5	7	3	3	1	1.54056	
B5	*P* $2_1/c$	14.80	11.68	8.48	90	93.68	90	1461.84	1	7	3	3	1	1.54056	
B6	*P* $2_1/c$	12.83	13.06	13.83	90	100.48	90	687.72	1	7	3	3	1	1.54056	
B7	*P* $2_1/n$	7.54	11.16	13.91	90	92.86	90	1168.30	1	7	3	3	1	1.54056	
B8	*P* 2_1	7.52	8.62	9.74	90	110.36	90	592.15	1	8	3	3	2	1.54056	
B9	*P* $2_1/n$	7.09	9.23	11.62	90	97.82	90	753.94	1	8	3	3	2	1.54056	
B10	*P b c a*	17.17	11.78	7.21	90	90	90	1458.02	1	8	3	3	2	1.54056	[001]
B11	*P* $2_1/c$	10.20	4.96	14.44	90	99.17	90	720.67	1	8	3	3	2	1.54056	
B12	*P* $\bar{1}$	10.60	11.60	5.50	97.88	103.89	71.46	621.43	1	9	3	3	3	1.79000	[121]

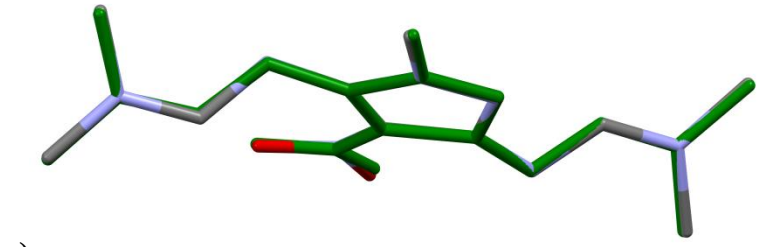
No	Space Group	a (Å)	b (Å)	c (Å)	α (^{0})	β (^{0})	γ (^{0})	Volume (Å^3)	Z'	Total DoF	DoF Position	DoF Orient	DOF Torsion	λ (Å)	PO
B13	*P* $2_1/c$	8.06	8.81	16	90	99.45	90	1120.66	1	9	3	3	3	1.54060	
B14	*P* $2_1/n$	21.74	10.06	9.45	90	99.96	90	2034.72	1	9	3	3	3	0.80098	
B15	*P* $2_1/c$	12.62	8.91	17.27	90	102.85	90	1894.18	1	9	3	3	3	1.54060	[100]
B16	*P b c a*	36.63	12.51	7.58	90	90	90	3470.96	1	9	3	3	3	1.54059	
B17	*P* $2_1/c$	11.58	8.09	17.22	90	107.20	90	1541.82	0.5	10	3	3	4	1.54056	
B18	*C* 2/*c*	19.73	4.79	19.25	90	108.96	90	1720.51	0.5	10	3	3	4	1.79000	
B19	*P* $2_1/n$	26.82	7.76	6.08	90	94.03	90	1261.82	1	10	3	3	4	1.54056	
B20	*P* $2_1/n$	13.39	5.14	14.66	90	112.65	90	931.81	1	10	3	3	4	1.54060	
B21	*P* $2_1/c$	19.04	17.43	17.42	90	113.82	90	5287.66	1	10	3	3	4	0.51966	[100]
B22	*P* $\bar{1}$	7.94	10.46	11.47	111.08	74.56	92.25	855.04	0.5	11	3	3	5	1.54056	
B23	*P n a* 2_1	6.12	7.47	32.99	90	90	90	1507.84	1	11	3	3	5	1.54056	
B24	*P* 2_1	10.15	8.70	13.69	90	110.65	90	1130.59	1	11	3	3	5	1.54056	[010]
B25	*P* $2_12_12_1$	10.93	12.72	15.71	90	90	90	2183.29	1	11	6	3	2	0.69200	
B26	*P* $2_12_12_1$	5.95	15.89	21.38	90	90	90	2019.69	1	12	3	3	6	1.54060	
B27	*P* $2_1/c$	7.43	13.31	14.03	90	95.32	90	1379.94	1	12	6	3	3	1.54056	
B28	*P* $2_1/c$	7.67	10.55	18.89	90	95.58	90	1521.00	2	12	6	6	0	1.54056	
B29	*P* $2_1/a$	13.23	5.11	19.66	90	107.67	90	1267.06	0.5	13	3	3	7	1.54056	
B30	*P* $\bar{1}$	9.58	9.97	7.60	106.11	95.12	78.22	682.40	1	13	3	3	7	1.54056	[511]
B31	*P* $2_1/c$	18.78	18.10	8.01	90	97.06	90	2701.25	1	13	3	3	7	1.14981	
B32	*P* $2_1/a$	16.06	13.09	13.32	90	99.40	90	2764.21	1	13	3	3	7	1.14981	
B33	*P* $2_1/c$	15.51	8.61	14.03	90	96.69	90	1859.40	1	13	6	3	4	1.54056	
B34	*P* $2_12_12_1$	14.45	34.69	8.711	90	90	90	4367.52	1	14	3	3	8	1.30000	[010]
B35	*P* $\bar{1}$	7.27	10.32	12.18	96.46	95.87	109.85	843.78	1	14	3	3	8	1.54056	[100]
B36	*P* $\bar{1}$	5.13	9.23	17.41	95.86	95.51	91.80	815.42	1	14	3	3	8	1.54056	[100]
B37	*P* $2_1/c$	14.58	8.54	13.78	90	96.07	90	1707.63	1	14	3	3	8	1.54056	[010]
B38	*P* $2_1/c$	14.72	5.99	20.79	90	114.82	90	1662.32	1	14	3	3	8	1.54056	
B39	*P* 2_1	11.72	6.83	11.05	90	94.38	90	881.67	1	14	3	3	8	1.54056	[010]

No	Space Group	a (Å)	b (Å)	c (Å)	α (^{0})	β (^{0})	γ (^{0})	Volume (Å^3)	Z'	Total DoF	DoF Position	DoF Orient	DOF Torsion	λ (Å)	PO
B40	$P\,2_1/c$	8.68	18.56	12.10	90	90.38	90	1948.06	1	14	3	3	8	1.54056	[100]
B41	$P\,\bar{1}$	7.98	8.88	11.10	86.69	75.08	73.20	728.41	1	14	6	6	2	1.54056	
B42	$P\,2_12_12_1$	30.03	11.23	5.89	90	90	90	1987.09	1	14	6	3	5	1.54056	[100]
B43	$P\,2_12_12_1$	22.79	13.02	6.920	90	90	90	2052.85	1	16	6	3	7	1.54056	[001]
B44	$P\,2_1/c$	5.25	12.24	21.41	90	93.24	90	1374.05	1	16	6	3	7	1.54056	
B45	$P\,2_12_12_1$	6.06	16.78	16.89	90	90	90	1717.80	1	16	6	6	4	1.54056	
B46	$P\,2_1$	11.88	5.98	13.08	90	113.51	90	851.42	1	16	6	6	4	1.54056	
B47	$P\,\bar{1}$	7.40	8.57	13.69	106.21	90.85	98.78	822.26	1	18	6	3	9	1.00045	[001]
B48	$P\,1$	7.63	19.66	5.06	95.65	105.43	81.00	721.01	2	20	6	6	8	1.54056	
B49	$P\,2_12_12_1$	35.94	12.92	5.00	90	90	90	2319.37	1	20	3	3	14	1.54056	[100]
B50	$P\,\bar{1}$	5.69	10.59	18.53	73.32	87.84	76.13	1037.86	0.5	21	12	3	6	1.54056	
B51	$P\,\bar{1}$	5.69	10.61	18.56	72.83	88.27	76.42	1039.37	0.5	21	12	3	6	1.54056	
B52	$P\,\bar{1}$	6.01	10.82	18.09	85.68	86.39	75.78	1136.55	1	24	15	3	6	0.64980	
B53	$P\,2_1$	14.55	5.90	14.24	90	112.83	90	1124.84	1	25	9	3	13	1.54056	
B54	$I\,2$	21.03	9.11	24.38	90	98.34	90	4622.43	2	26	6	6	14	1.00140	
B55	$P\,\bar{1}$	5.65	19.56	22.07	84.22	88.41	83.60	2411.72	4	28	12	12	4	1.54056	
B56	$P\,\bar{1}$	11.55	16.65	7.86	95.40	94.64	71.25	1424.06	1	28	9	9	10	1.54059	[001]
B57	$P\,2_12_12_1$	13.44	50.29	27.06	90	103.15	90	1872.12	1	28	3	3	22	1.54056	
B58	$P\,2_1/c$	9.55	22.29	22.07	90	93.75	90	4686.28	2	30	12	6	12	0.79977	
B59	$P\,2_12_12_1$	24.30	20.57	4.87	90	90	90	2433.30	3	33	9	9	15	0.49957	
B60	$P\,2_1/c$	12.36	8.56	37.30	90	92.88	90	3942.30	2	42	18	18	6	1.54056	
B61	$P\,\bar{1}$	11.67	56.51	5.43	73.06	100.02	120.08	301.82	1	49	3	3	43	0.85005	

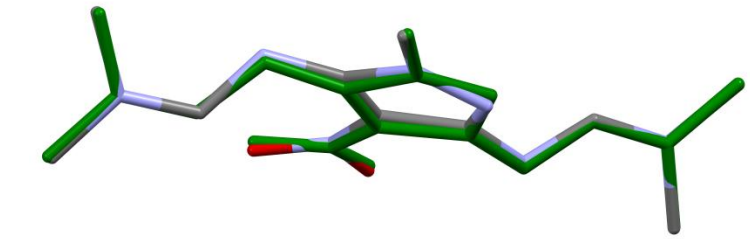
3.3 Results – the baseline DASH performance

A summary of the results from all SA runs as described in the experimental section, including information on the Pawley fit, is given in Table 3.2 and Table 3.3.

It is worth noting that not all solutions reported have the same χ^2 values. Whilst there is only one χ^2 value corresponding to the global minimum, there is a range of χ^2 values which correspond to crystal structures which are so close to the true crystal structure, that they were legitimately considered successful (*i.e*. could easily be refined to the published model). By way of an example, the nitropyrazole (B30) returned a best profile χ^2 of 10.49, but solutions with values of up to 15.21 were considered successful (Figure 3.2).



a)



b)

Figure 3.2 The crystal structure overlay of the CSD deposited crystal structure (in green) and a) the best DASH solution (χ^2 = 10.49); and b) the DASH solution with χ^2 = 15.21. The overlay of only one of the fifteen molecules is presented for simplicity. Hydrogen atoms have been omitted. The RMSD values achieved with the 'crystal packing similarity' feature in Mogul, are 0.065 Å and 0.115Å, respectively, which gave the confidence to consider (b) as a successful DASH solution.

Table 3.2 A summary of the Pawley refinement details and baseline DASH performance against the FDS based on the 50 and 100 SA runs (1×10^7 moves). † indicates cases where 100 SA runs were required

No	Total DoF	Resolution (Å)	No. of Reflections	Pawley χ^2	Best Profile χ^2	χ^2 ratio	Success Rate (%)	RMSD (Å)
A1	6	1.17	389	13.85	28.41	2.05	100	0.034
A2	7	1.75	136	3.88	9.35	2.41	100	0.024
A3	8	1.57	214	3.68	6.27	1.70	100	0.021
A4	9	1.83	153	8.45	32.67	3.87	100	NA
A5	10	1.66	168	3.26	8.41	2.58	100	0.029
A6	10	2.03	96	3.41	9.10	2.67	100	NA
A7	10	2.10	143	1.99	3.18	1.60	48	0.032
A8	11	1.75	208	2.63	4.77	1.81	100	0.037
A9	11	1.76	263	4.40	30.09	6.84	100	0.083
A10	11	1.82	103	8.20	62.01	7.56	100	0.119
A11	12	1.86	148	8.52	23.92	2.81	100	0.0541
A12	12	1.52	219	5.57	22.92	4.11	100	0.059
A13	12	2.20	124	4.21	35.89	8.52	78	0.155
A14	13	1.85	140	3.21	18.51	5.77	96	0.066
A15	13	1.77	318	3.48	10.57	3.04	100	0.120
A16	13	1.54	228	8.89	22.91	2.58	42	0.147
A17	13	2.08	164	12.16	95.34	7.84	100	0.141
A18	14	2.29	148	3.40	13.44	3.95	4	0.645
A19	14	1.86	499	0.56	2.75	4.91	14	0.104
A20	15	1.86	228	5.25	11.44	2.18	34	0.095
A21	15	2.13	174	2.01	4.40	2.19	56	0.022
A22	16	3.18	72	2.71	11.23	4.14	28	0.085
A23	16	2.19	161	5.23	14.35	2.74	54	0.087
A24	16	1.81	336	3.70	12.79	3.46	50	0.136
A25	17	1.76	338	7.87	38.05†	4.83	2†	0.266
A26	17	2.62	126	2.14	15.29†	7.14	1†	0.099
A27	17	2.00	369	3.39	23.70	6.99	78	0.053
A28	18	1.63	341	3.81	18.04	4.73	8	0.139
A29	18	2.37	182	2.41	12.51	5.19	60	0.045
A30	18	2.31	252	0.81	10.65	13.15	34	0.376
A31	19	1.97	262	4.07	11.37	2.79	16	0.081
A32	20	1.68	320	3.63	9.47	2.61	18	0.048
A33	20	1.99	180	29.09	144.57	4.97	14	0.127
A34	22	1.76	518	4.04	10.90	2.70	4	0.087
A35	24	2.65	123	2.68	8.25	3.08	14	0.126
A36	24	3.13	111	3.44	5.89	1.71	46	0.180
A37	26	1.80	567	0.34	4.86	14.29	0	NA
A38	28	2.80	218	3.47	7.81	2.25	98	0.118
A39	28	2.92	195	7.93	107.29†	13.53	1†	0.263
A40	30	2.04	362	11.32	207.38†	18.32	0†	NA
B1	6	1.67	76	6.48	15.02	2.32	92	0.070
B2	6	1.41	165	0.50	4.37	8.74	100	0.067
B3	6	3.64	19	1.45	2.67	1.84	100	0.281
B4	7	1.68	120	4.56	9.02	1.98	100	0.031
B5	7	2.60	87	4.37	36.07	8.25	100	0.117
B6	7	1.44	243	2.76	5.28	1.91	100	0.031
B7	7	1.64	276	3.74	9.16	2.45	100	0.011
B8	8	1.98	94	5.15	16.4	3.18	100	0.098
B9	8	1.44	267	3.48	8.62	2.48	100	0.123
B10	8	1.52	223	5.87	18.4	3.13	100	0.130

No	Total DoF	Resolution (Å)	No. of Reflections	Pawley χ^2	Best Profile χ^2	χ^2 ratio	Success Rate (%)	RMSD (Å)
B11	8	1.62	173	11.00	26.03	2.37	100	0.083
B12	9	1.74	245	6.47	19.65	18.05	96	0.118
B13	9	1.52	331	2.01	67.14	33.40	100	0.109
B14	9	1.86	327	5.75	12.27	2.13	100	0.012
B15	9	2.13	204	6.93	20.78	3.00	66	0.045
B16	9	2.49	113	4.34	13.03	3.00	100	0.042
B17	10	1.80	288	5.61	14.20	2.53	100	0.079
B18	10	1.82	154	1.11	2.13	1.92	70	0.102
B19	10	2.055	148	1.70	3.92	2.31	100	0.188
B20	10	2.25	87	8.60	46.94	5.46	100	0.264
B21	10	2.03	280	0.47	0.84	1.79	44	0.331
B22	11	1.90	264	5.45	12.62	2.32	100	0.057
B23	11	2.04	98	1.60	3.64	2.28	98	NA
B24	11	1.86	214	4.44	19.69	4.43	96	0.217
B25	11	2.01	188	4.62	17.83	3.86	100	0.075
B26	12	2.24	133	1.43	3.18	2.22	84	0.014
B27	12	1.97	188	5.77	10.34	1.79	44	0.026
B28	12	2.17	156	1.07	3.93	3.67	100	0.075
B29	13	1.83	220	7.99	16.87	2.11	92	0.171
B30	13	2.06	165	4.89	10.49	2.15	64	0.065
B31	13	2.22	260	1.49	3.15	2.11	58	0.160
B32	13	2.60	160	4.88	10.69	2.19	100	0.142
B33	13	1.85	306	5.18	16.55	3.19	100	0.067
B34	14	1.90	427	28.99	47.97	1.65	50	0.052
B35	14	2.64	95	5.57	13.25	2.38	14	0.182
B36	14	2.40	123	4.79	24.17	5.05	4	0.060
B37	14	2.13	184	2.83	7.05	2.49	12	0.039
B38	14	2.66	93	2.43	15.26	6.28	36	0.233
B39	14	2.17	111	87.65	149.48	1.71	4	0.175
B40	14	2.24	184	63.03	124.28	1.97	8	0.126
B41	14	2.16	150	1.00	2.24	2.24	98	0.765
B42	14	2.52	100	229.92	1252.69	5.45	20	0.280
B43	16	2.22	135	55.26	220.77	4.00	12	0.158
B44	16	1.76	256	2.00	13.05	6.53	8	0.103
B45	16	1.55	311	8.58	13.81	1.61	14	0.079
B46	16	1.55	289	3.16	5.61	1.78	4	0.068
B47	18	2.53	103	27.61	61.01	2.21	14	0.017
B48	20	2.39	110	9.20	36.09†	3.92	4†	0.130
B49	20	1.86	260	52.10	373.79†	7.17	0†	NA
B50	21	2.87	88	1.25	60.04†	48.03	0†	NA
B51	21	2.59	121	4.04	37.71†	9.33	1†	0.197
B52	24	2.03	280	1.09	7.36†	6.75	0†	NA
B53	25	1.75	237	3.51	13.50	3.85	2	0.077
B54	26	2.32	230	0.04	0.27†	6.75	0†	NA
B55	28	2.92	196	3.79	25.29	6.67	4	0.081
B56	28	2.71	148	43.42	381.91†	8.80	0†	NA
B57	28	2.17	257	4.69	221.32†	47.18	0†	NA
B58	30	2.62	276	9.09	29.26	3.22	78	NA
B59	33	1.67	358	16.74	523.34†	31.26	0†	NA
B60	42	2.45	280	1.77	26.75†	13.53	0†	NA
B61	49	2.64	332	22.52	602.88†	18.32	0†	NA

Table 3.3 A summary of the Pawley refinement details and baseline DASH performance against the FDS based on the 500 SA runs (5×10^7 moves).

No	Total DoF	Resolution	No Reflections	Pawley χ^2	Best Profile χ^2	χ^2 ratio	Success Rate (%)	RMSD (Å)
A37	26	1.80	567	0.34	4.86	14.29	0	NA
A40	30	2.04	362	11.32	74.18	6.55	0.2	0.296
B49	20	1.86	260	52.10	366.32	7.03	0	NA
B50	21	2.87	88	1.25	15.24	12.19	0.2	0.488
B52	24	2.03	280	1.09	1.88	1.72	9.4	0.082
B54	26	2.32	230	0.04	0.09	2.25	2	0.013
B56	28	2.71	148	43.42	210.68	4.85	0	NA
B57	28	2.17	257	4.69	170.71	36.40	0	NA
B59	33	1.67	358	16.74	266.84	15.94	0	NA
B60	42	2.45	280	1.77	5.94	3.36	0.4	0.498
B61	49	2.64	332	22.52	509.73	22.63	0	NA

3.4 Discussion

3.4.1 Dataset analysis

The distribution of structures within the FDS, with regard to space group and DoF, was compared to that of small organic powder crystal structures (deposited at the CSD), to demonstrate that it was truly representative of the current PXRD landscape.

3.4.1.1 Space Group trends

It is well established that 80% of molecular organic compounds crystallise in one of the following five space groups: $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$ and $C2/c$ (Brock and Dunitz, 1994; Srinivasan, 1991). The FDS is broadly representative of the space group distribution in the CSD. Table 3.4 and Figure 3.3 show a good agreement between the population of space groups of organic structures in the CSD determined by PXRD and that in the FDS.

Table 3.4 Distribution of space groups within the FDS and the CSD. *the named space group or equivalent

Space Group*	No. of structures in FDS	No. of organic powder structures in CSD	% of structures in FDS	% of organic powder structures in CSD
$P\,2_1/c$	40	355	39.6	35.8
$P\,\bar{1}$	23	185	22.8	18.6
$P\,2_12_12_1$	16	146	15.8	14.7
$P\,2_1$	10	100	9.9	10.1
$P\,b\,c\,a$	3	63	3.0	6.4
$P\,n\,a\,2_1$	2	33	2.0	3.3
$C\,2/c$	2	41	2.0	4.1
$P\,1$	1	22	1.0	2.2
$P\,b\,c\,2_1$	1	20	1.0	2.0
$I\,2$	1	13	1.0	1.3
$P\,b\,c\,n$	1	7	1.0	0.7
$C\,m\,c\,a$	1	7	1.0	0.7

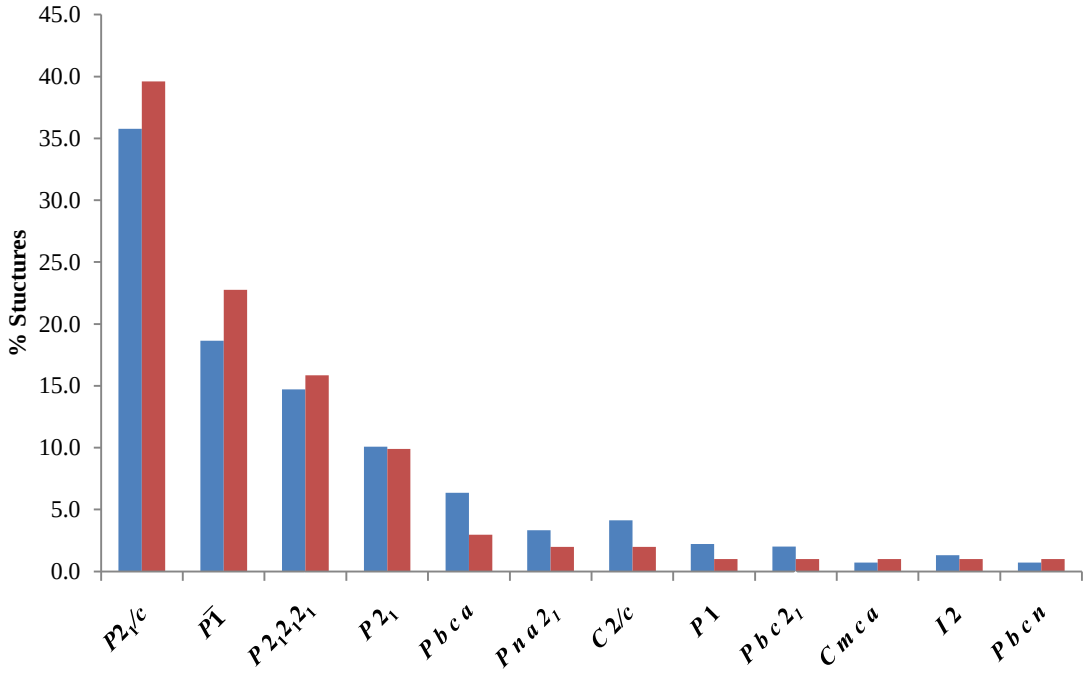


Figure 3.3 A comparison of the relative space group distribution between the FDS (red) and the CSD (blue)

3.4.1.2 DoF trends

With increasing DoF, a pseudo-exponential decay of the number of crystal structures solved from powder data is observed. The distribution of the structural complexity within the FDS differed, with just over 50% of the structures having DoF ≥ 14, in order to satisfy the aim of improving performance in this upper range of structural complexity (Figure 3.4).

During the search for relevant crystal structures and their associated diffraction data, it was observed that the majority of IUCr published crystal structures (solved from powder data) had

fewer than 15 DoF. Unsurprisingly, this appeared to be the general trend in the CSD, with approximately 70 % of the deposited organic powder crystal structures having up to 15 DoF and only just over 6% of the structures possessing more than 30 DoF. Examination of those 6% revealed these structures to be 'unusual' (*e.g.* peptides, glycerols, acylglycerols, etc.), falling outside the range of interest of this work and therefore not eligible for inclusion in the FDS. However, for completeness, a single example of this complexity type was added: the 1,2,3,-tris(nonadecanoyl)glycerol (polymorph β) (B61), with the expectation that this would be an example of an intractable crystal structure. It is worth noting that whilst this example was previously solved from powder data, the starting conformation was derived from the crystal structure of β-1,2,3-tris(octodecanoyl)glycerol (van Langevelde *et al.*, 2001; Van Langevelde *et al.*, 2000) and the addition of a CH_3 group at the end of each chain. Thus the problem was reduced to one of just finding the position and orientation of the molecule in the unit cell and was (as reported) only a 6 DoF problem rather than the 49 DoF problem considered here.

3.4.1.3 Additional remarks

The resolution (minimum *d*-spacing) of the powder data and the number of reflections used in the Pawley refinement are two fundamental factors expected to influence both the SR and the quality of the DASH solution.

Large variations of those factors were observed within the FDS, with B3 having the lowest resolution (only 3.64 Å) and lowest number of reflections (only 19). The powder X-ray diffraction data of B3 (Figure 3.5), which was deposited by Schmidt *et al.* (2005), exhibits very broad peaks and was only collected to 34˚ 2θ (resolution 2.6 Å) as its main purpose was to serve as a reference for crystal structure prediction calculations. It is important to note that the DASH runs were performed using the data to only 3.64 Å because of an inability to obtain a satisfactory Pawley fit to any higher resolution using DASH. Whilst not usually a significant restriction on the ability of DASH to deal with a powder dataset, it is undoubtedly a limitation in some cases and reflects a need for some improvements in the core least-squares fitting routines in the program.

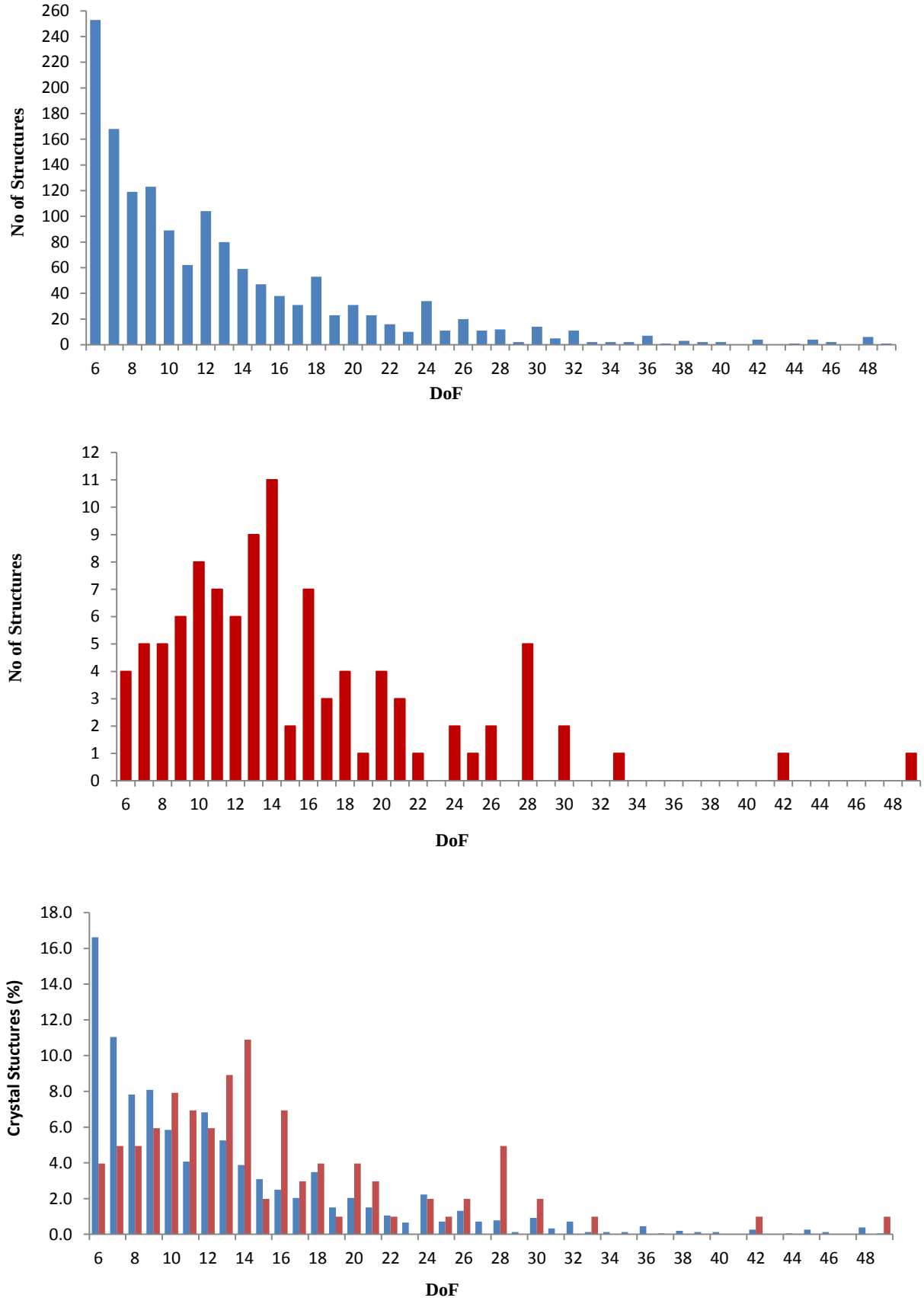


Figure 3.4 The distribution of crystal structures plotted as a function of their DoF. The upper plot is based on the organic powder crystal structures in the CSD; the middle plot corresponds to the distribution within the FDS; and the lower plot is and overlay of the upper and middle graphs, represented as their relative distributions.

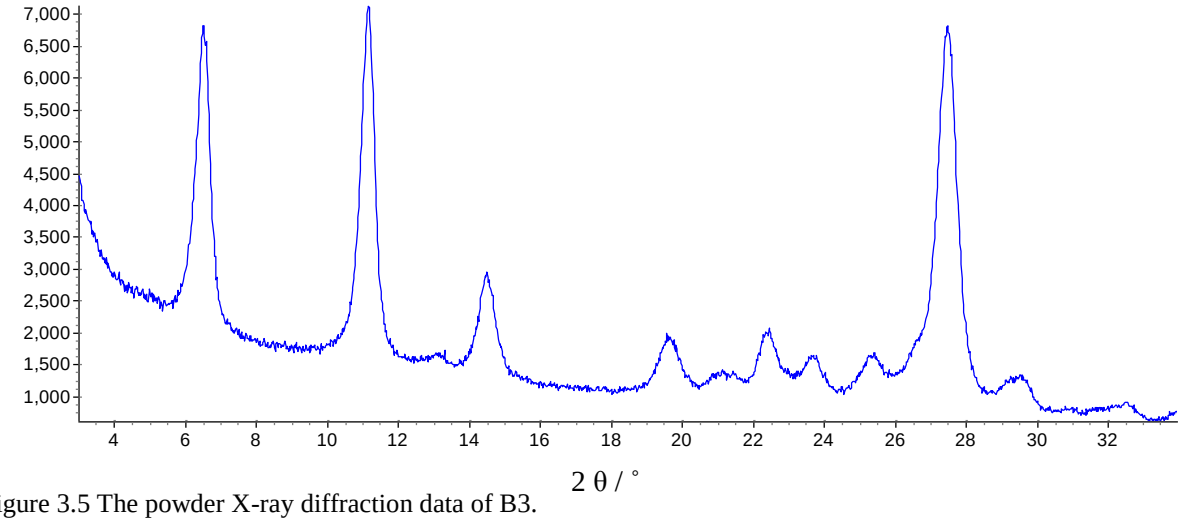


Figure 3.5 The powder X-ray diffraction data of B3.

Due to the simplicity of B3 (only 6 DoF), its observed SR was high, but the quality of the solutions was affected, resulting in a best RMSD value of 0.281 Å. The low solution quality is clearly demonstrated in Figure 3.6 where the solved structure is visibly an offset from the true solution. However, this DASH solution was considered to be successful as the offset could be addressed by Rietveld refinement.

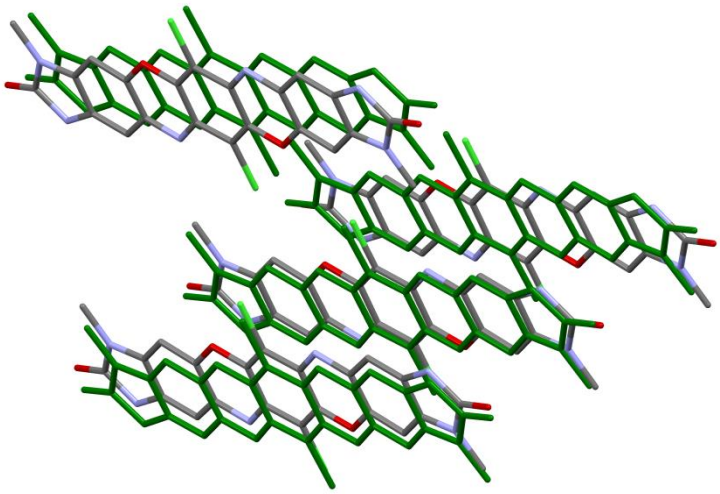


Figure 3.6 Overlay of the best DASH solution of B3 and the CSD deposited crystal structure (in green; CSD reference code QAMQOL). The initially minor positional shift of the molecule progressively worsens with the application of the symmetry operations. The overlay of only four molecules (rather than the default 15) is presented for clarity. Hydrogen atoms have been omitted.

On the other hand, A1 had the highest resolution of 1.17 Å and A34 the highest number of reflections (518). A1 also provided a high success rate (100%), due to the simplicity of the crystal structure. However, in this case the quality of the structure solution is also very good, with an RMSD of only 0.034 Å for the 15 molecules overlay.

Shankland *et al*. (2002) noted that in the case of famotidine, good quality solutions could normally be found when the data resolution was better than 2.5 Å and none of the results obtained in this work with the FDS contradict this finding. It is clear from equation 1.3 that the structure factor calculation time will increase linearly with the number of reflections to be calculated. Whilst it may therefore be tempting to reduce the number of reflections in order to speed up the evaluation of each trial crystal structure, it is recommended that the above 2.5 Å resolution 'rule' is adhered to.

3.4.2 Baseline DASH performance

A graphical representation of the baseline SR obtained using DASH with its SA control parameters set to their default values is shown in Figure 3.7. The better population of the SR *vs* DoF landscape shows that the three complexity groups established by Florence[7] *et al*. are still very relevant. However, the addition of a further, fourth group can be considered, resulting in the following categories:

1) simple crystal structures with DoF < 14

2) moderately complex crystal structures, with 14 ≥ DoF ≤ 20

3) complex crystal structures, with 21 ≥ DoF ≤ 30

4) intractable crystal structures with DoF > 30.

It is of course true that these groups remain only a *broad* description of crystal structure complexity and as such exceptions will be observed; *e.g*. apparently simple structures that prove difficult to solve or complex structures that solve more easily than expected.

[7] Florence *et al*. divided crystal structures into: 1) simple with DoF <15; 2) moderate 15 ≥ DoF ≤ 20; and 3) complex with DoF >20

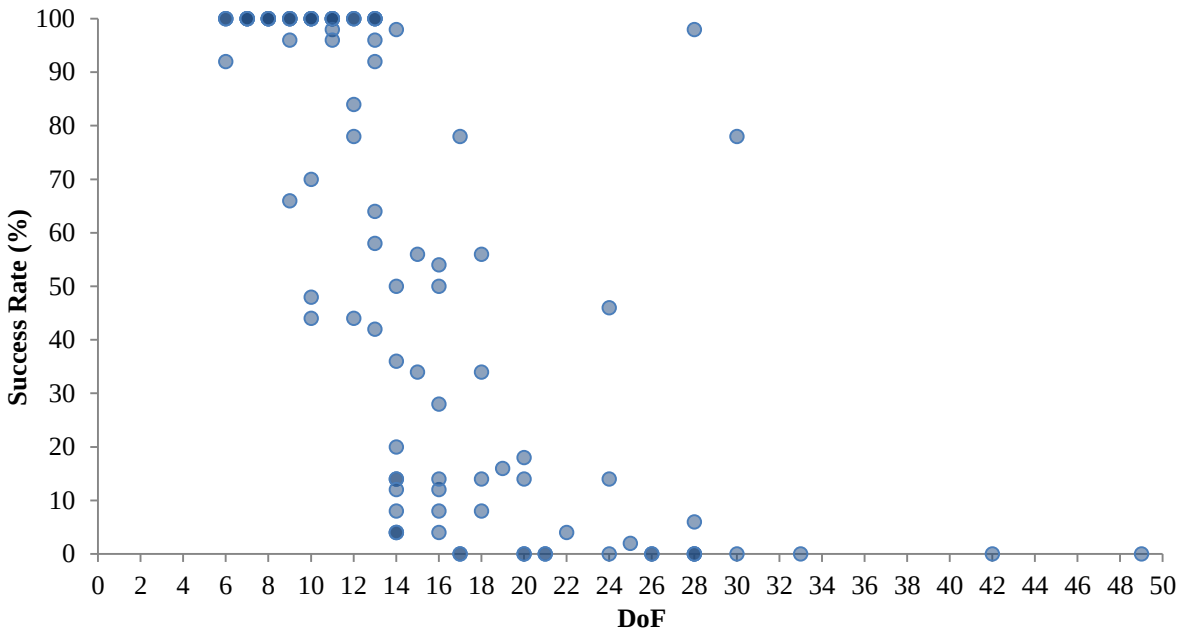


Figure 3.7 Graphical representation of the baseline DASH success rate as a function of the total degrees of freedom (SA parameters used: T_0=0; CR = 0.02; N_1=20; and N_2=25). Again, the depth of shading of each point is proportional to the number of structures representative of a particular DoF.

Importantly, the results reported here for the baseline DASH performance against the original 35 structures studied by Florence *et al.*, were very close to those reported in the original publication. In particular, for the simple crystal structures practically identical frequencies of success were observed. With compounds of higher complexity, small differences in the reported SRs are observed and these differences can be attributed to three main factors:

1) Only 20 SA runs were carried out by Florence *et al*. for each compound. In this work, up to 100 SA runs were employed. As a larger number of runs is more likely to obtain a better estimate of the true SA success rate, the results of this work can be taken to be more accurate in this regard.

2) Whilst reasonable efforts were made to closely replicate the 'conditions' employed by Florence *et al.*, minor differences in the Pawley fitting results (such as the number of the reflections, the unit cell parameters, the final profile χ^2) change the nature of the hypersurface being explored by the SA.

3) Finally, subjective and objective differences in the criteria used to decide whether or not a structure solution has truly been obtained also play a role. A particular example of this is found for A16:

Florence et al; Pawley χ^2=7.67, reported successful χ^2 up to 72.45

This work: Pawley χ^2=8.89, successful χ^2 up to 24.2

In general, the criteria used in this work can be considered to be more stringent, reflecting our ability to assess the agreement between a putative solution and its target (using Mercury's structure overlay feature) and the ease with which verification Rietveld refinements can be carried out.

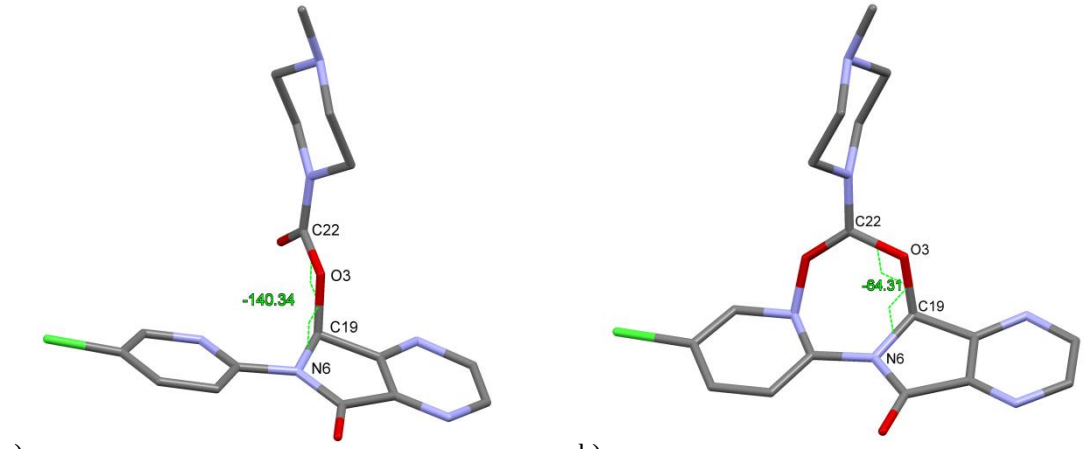
3.4.2.1 Simple crystal structures, DoF <14

Each DoF in this group of structural complexity has an average SR close to 100%. As the group with the most representatives (exactly 50 crystal structures), the conclusion that simple crystal structures solve easily and with a good reproducibility can be considered to be a reliable one. All of the structure solutions obtained were in excellent agreement with the reference crystal structures. Furthermore, these solutions were found with considerably fewer SA moves than the maximum 10^7 allowed (results not shown[8]) and so the 10^7 SA moves can be considered more than sufficient to maintain a high SR for this complexity group. In practice, when the objective is simply to solve the crystal structure, 10^7 moves can be considered excessive and 5 x 10^6 is much more appropriate.

One significant change from the work of Florence *et al*. is the appearance of simple examples with significantly reduced SRs. A number of the FDS compounds reported here (*e.g*. A7, A16, B21 and B27), exhibit markedly lower SRs of between 40% and 50%.

In the case of A7, the SR reduction can be rationalised in terms of a strong tendency for the SA to get trapped in a specific local minimum that happens to bear a reasonable resemblance to the true crystal structure (see Figure 3.8). For A16, B21 and B27 (all of which have acceptable resolution [1.54-2.1Å] and number of reflections [143-280]) it must also be the case that there are local minima from which the SA has trouble escaping, though these minima do not have any obvious structural correlation with the correct crystal structure.

[8] DASH allows one to 'visualise' the progress of an SA run by outputting the best χ^2 against the number of SA moves. By plotting χ^2 against the number of moves for multiple SA runs, one can clearly see the point at which a valid solution is normally obtained (*i.e*. close to Pawley χ^2) and so obtain a better estimate of the number of SA moves actually needed to solve the structure.



a) b)

Figure 3.8 The crystal structure conformation of A7: a) the best DASH solution, $\chi^2 = 3.18$; and b) the next best solution, representing a local minimum, $\chi^2 = 19.16$. The formation of a pseudo-7-membered ring[9] is observed in (b) due to the rotation of the N6-C19-O3-C22 torsion angle. Judged by the number of occurrences of (b) in the final DASH results (close to 50%) this local minimum proves difficult to escape from. The hydrogen atoms have been omitted for clarity.

3.4.2.2 Crystal structures of moderate complexity, $14 \geq \text{DoF} \leq 20$

A number of compounds were selected for this complexity group in pursuit of a well-defined trend in the SR reduction with respect to complexity. Compounds with 14–16 DoF were of particular interest, as the results reported by Florence *et al*. showed a SR drop between 14 and 15 DoF.

With the larger number of representatives in this complexity group, it is fair to say that there is no clear trend in the SR reduction. Indeed, large variations of SR are observed, with a maximum SR of 98% (B41, 14 DoF) and a minimum SR of 0% (B49, 20 DoF) following the initial set of 100 SA runs, 1×10^7 SA moves.

The feature that differentiates this group from the previously reported group of moderate complexity is the fact that it starts at 14 DoF (*c.f*. 15 in the work of Florence *et al*.) and it shows clearly that there is a sharp drop in SR above 13 DoF. Only 1 of the 11 compounds with DoF = 14 exhibits a SR > 60%, compared with the 7 out of 9 structures with DoF = 13. The trend can also be illustrated by considering the average SR for each DoF (Table 3.5).

[9] From DASH's stand point, there is no bond formation; rather the atoms are close enough for the formation of a chemical bond. More importantly, the N6-C19-O3-C22 torsion is still allowed to rotate during the optimisation, providing that the SA can escape this particular local minimum.

Table 3.5 The Average SR of each DoF between 6 and 20. Note that the results are based on the DASH calculations of up to 100SA runs.

DoF	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of representatives	4	5	5	6	8	7	6	9	11	2	7	3	4	1	4
Average SR (%)	98	100	100	94	83	99	84	84	25	48	27	26	31	16	13

Overall, moderately complex crystal structures have a lower, but reasonable SR, with a group average of approximately 26%.

The quality of the solutions, when compared to the reference structures, remains high. An apparent exception worth noting is the compound B41, for which the RMSD value is 0.765 Å. Although such a high RMSD value would normally indicate a very poor correspondence with the deposited structure, close inspection of the 15 molecule overlay (Figure 3.9) shows that the high RMSD is due to only one incorrect torsion angle (C_4-C_8-S_2-N_2) in the best DASH solution. The electron density is still satisfied as the local minimum is merely a 120° rotation of C_4-C_8-S_2-N_2 from being correct; the 'swapping' of nitrogen and oxygen atom types has only a negligible effect upon the SA χ^2. Such rotational errors are not uncommon in SDPD of molecules with SO_2NH_2 groups and can usually be identified (and corrected) by carefully considering hydrogen bonding in the packed crystal structure.

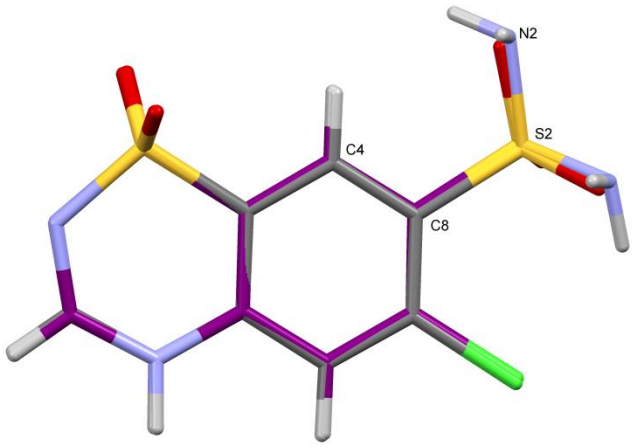


Figure 3.9 The crystal structure overlay of the best DASH solution for B41and its reference (the carbon atoms of which have been are given in purple for ease of comparison).

3.4.2.3 Complex crystal structures, 21 ≥ DoF ≤ 30

As previously indicated, the DoF given as boundaries for the various complexity groups are not rigid; rather they are broad guides, and this is clear with this complexity group. Whilst the lower boundary can be defined relatively easily due to the reduction in the average SR from

13% to 0.5%, for DoF 20 and 21 respectively (Table 3.6), the upper boundary is hard to define due to the number of crystal structures which remained unsolved after the 500 SA runs (see below).

Data resolution / number of reflections needs to be taken into account when discussing the low SR of the compounds of 21 DoF. For example, low SRs for B50 (2.87 Å, 88 reflections), and B51 (2.59 Å, 121 reflections) are to be expected.

Again, there is no clear trend in the relationship between SR and DoF in this group. Overall, SRs are low and A38 (DoF 28, SR 98%) and B58 (DoF 30, SR 78%) are clear outliers. Unfortunately, there is no clear explanation for why compounds of such complexity exhibited these high SRs.

Table 3.6 The Average SR of each DoF between 20 and 30. Note that the results are based on the DASH calculations of up to 100SA runs.

DoF	20	21	22	24	25	26	28	30	33	42	49
Number of representatives	4	2	1	3	1	2	5	2	1	1	1
Average SR (%)	13	0.5	2	30	0	3	20	39	0	0	0

The challenge presented by the compounds in this group is demonstrated by the need to perform 500 SA runs for 8 of the 19 compounds in order to find any solutions. Of those 8 compounds, 4 remained unsolved and further two gave a SR of only 0.2% , *i.e.* 1 in 500 runs was successful in reaching the crystal structure. For the remaining 2 crystal structures (of the 8), the extended number of SA step proved advantageous, and they achieved SRs of 2% and above.

Overall, the observed SR is very low and the reproducibility is poor, *i.e.* it is not unusual for the DASH solutions to have notable differences in their χ^2 values. Taking A34 as an example, the best DASH solution has a χ^2 of 11, whilst the χ^2 of the second best solution is 35. When the two solutions are overlaid, it becomes apparent that the position and orientation of the molecules in the unit cells are identical and the χ^2 difference is a result of the rotation of a single torsion angle (Figure 3.10). As such, the second top solution was considered refinable and therefore counted as successful.

Regardless of the low SR, the quality of the solutions, in general, remains high, as is demonstrated by the good RMSD values for 15 molecule overlays in Mercury (Table 3.2 and Table 3.3)

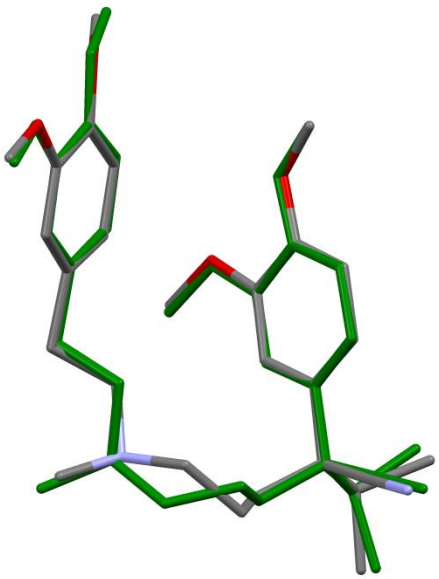


Figure 3.10 An overlay of the first (green) and second DASH solutions of A34 (verapamil HCl). For clarity only the overlay of one verapamil molecule from each crystal structure has been shown. The hydrogen atoms have been omitted for the same reason.

3.4.2.4 'Intractable' crystal structures, DoF > 30

This complexity group comprises only three representatives. This is largely a reflection of the fact that very few structures of this complexity are solved by powder diffraction. Nevertheless, such structures are of particular interest, as they fall into the range of structural complexity to which SPDP aspires.

None of the three compounds solved with the initial 100 SA runs, and when 500 runs were performed, only B60 reached a solution. The SR for B60, even with the extended number of SA moves (increased from 1×10^7 to 5×10^7) is less than one percent. Unsurprisingly, the best solution for B60 exhibited an incorrect SO_2NH_2 group rotation, as previously described for B41 (Figure 3.11), resulting in a poor RMSD even though the structure was largely correct.

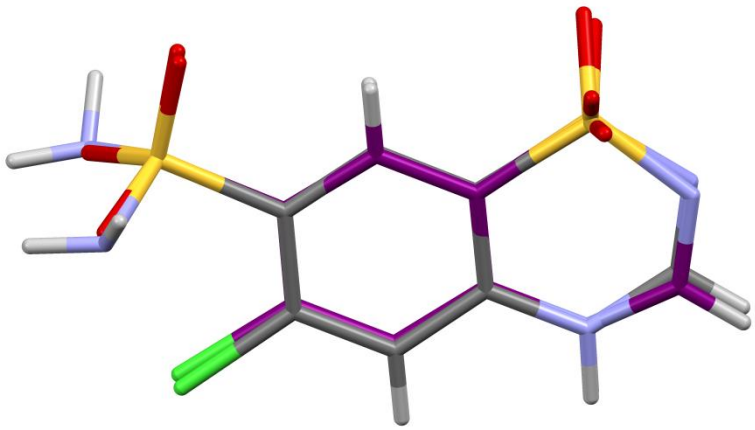


Figure 3.11 The crystal structure overlay of the best DASH solution for B60 and its reference (the carbon atoms of which have been are given in purple for ease of comparison).

When the composition of the DoF for these three compounds is considered (Table 3.7), it is notable that B60 has the fewest torsional DoF and the most positional / orientational DoF (a result of the six fragments in the asymmetric unit cell). This, combined with the fact that B60 solved, might suggest that the SA is more successful in cases where the majority of DoF are positional and orientational, rather than torsional. The validity of this supposition is investigated and further discussed below (Section 3.4.2.5, the statistical analysis).

Table 3.7 DoF composition of the intractable compounds. AU = asymmetric unit cell.

No.	Z'	No. fragments in AU	DoF_{total}	$DoF_{positional}$	$DoF_{orientational}$	$DoF_{torsional}$	Solved (500 SA runs)
B59	3	3	33	9	9	15	No
B60	2	6	42	18	18	6	Yes
B61	1	1	49	3	3	43	No

Unsurprisingly, B61 (43 optimisable torsion angles) failed to solve. Further analysis on the predicted minimum of SA runs required to achieve a solution for B61 (with no additional conformational information, in the form of constraints) is presented in Section 3.4.2.5.

3.4.2.5 Statistical analysis of the baseline DASH performance

The relatively simplistic categorisation of structural complexity outlined above is undoubtedly a useful guide as to the likely 'degree of difficulty' to be encountered during a structure determination. However, with 100 data sets, the possibility of a more statistically sound characterisation arises, which may allow better predictions to be made.

It is obvious from the approximate 's-shape' of the 'Success versus DoF' curve for the baseline DASH performance that simple linear regression is not appropriate. Consultation with statisticians[10] about the problem led to an analysis based on the empirical log-of-the-odds (ELO) transform. The ELO, as described by Cox and Snell (1989), takes the form given in equation 3.1

$$ELO = ln\left(\frac{r_i + 0.5}{n_i - r_i + 0.5}\right) \qquad \text{Equation 3.1}$$

where *i* is the subject (*i.e.* each of the individual compounds of the FDS), n_i is the maximum value of the sample (in this case the SR, *i.e.* 100%) and r_i is the error associated with it (*i.e.* the actual SR value). As such, Equation 3.1 can be re-written as Equation 3.2.

$$ELO = ln\left(\frac{SR_i + 0.5}{100 - SR_i + 0.5}\right) \qquad \text{Equation 3.2}$$

Initially, the ELO analysis was calculated based on the total DoF of the FDS's compounds in order to determine if there is any evidence of DoF_{total} influencing the resulting DASH performance in a predictable manner. The fit, performed using MINITAB, returned an R^2 of 53.73 and a p-value of 0 for the DoF_{total}, showing them to be a statistically significant factor in determining success rate.

$$ELO = ln((SR + 0.5)/(100 - SR + 0.5)) = 6.565 - 0.375 \times DoF_{total} \qquad \text{Equation 3.3}$$

Put simply, it can be concluded that the ELO transform of the SR reduces by 0.375 for each increase in DoF_{total}.

Using Equation 3.3, it is possible to calculate the likely success rate for any given problem based on its DoF_{total}. Taking A34 (DoF_{total} = 22) as an example, the predicted SR is given by

$$e^{(6.565-0.375\times22)} = 0.1854 = (SR_{predicted} + 0.5)/(100 - SR_{predicted} + 0.5)$$

$$SR_{predicted} = 0.1854 \times \frac{100}{1 + 0.854} \approx 16\%$$

[10] Statistical Advisory Service, Applied Statistics at the University of Reading

The actual SR for A34 is 4%. Following the same procedure, the predicted SR was calculated for all compounds of the FDS and the results are illustrated in Figure 3.12. The calculated fit is clearly over-optimistic but nevertheless models the general trend in success rate drop off.

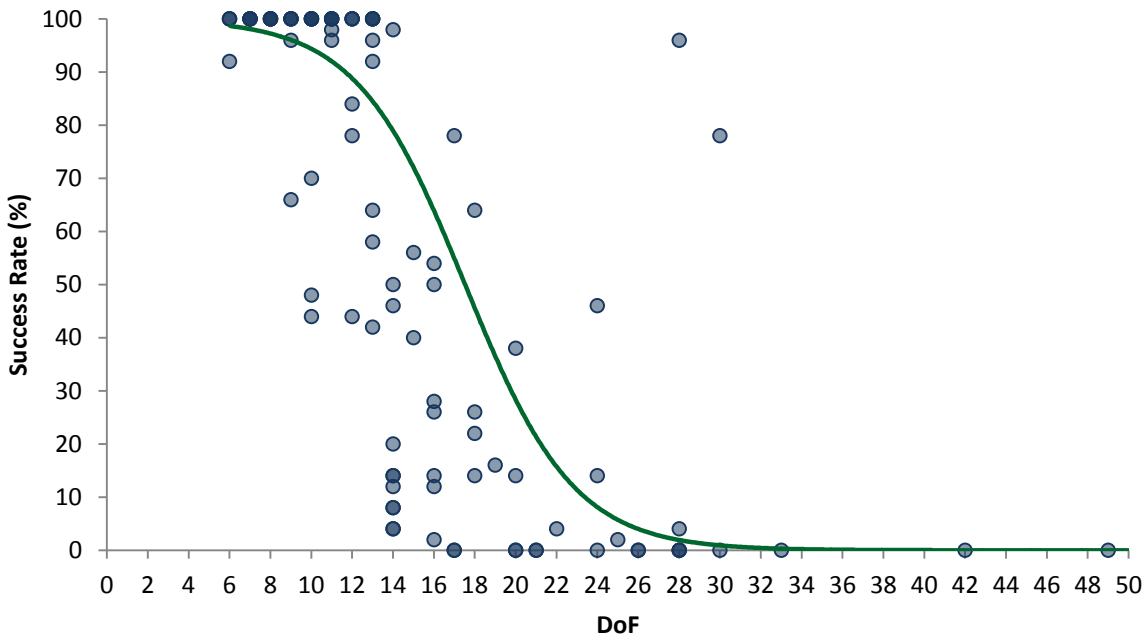


Figure 3.12 The ELO model based on the total DoF (solid green line). The observed SRs are shown in blue with increasing opacity with increased number of examples.

A more accurate model was sought, which took into account the various contributions to the DoF_{total}. Regression analysis of the ELO transform versus all of these components indicated that the orientational DoF of the problem have no significant impact on the observed SRs (Table 3.8) and so can be safely ignored in the subsequent analyses. It is important to note, however, that this conclusion is based only on the structures in the FDS, which do not exhibit large variations in their $DoF_{orientation}$[11].

Table 3.8 Regression analysis of the ELO *vs.* positional, orientational and torsional DoF.

Term	F-value	p-value
Regression	39.81	0
Positional DoF	16.14	0
Torsional DoF	77.76	0
Orientational DoF	1.69	0.197

[11] Single atom counterions (such as chloride) contribute only $DoF_{positional}$; equally the water molecules were represented as single oxygen atoms and as such contribute only $DoF_{positional}$.

Nevertheless, a final ELO transform was calculated taking into account only the positional and torsional DoF, resulting in the relationship given in Equation 3.4 (R^2 = 55.18).

$$ELO = ln((SR + 0.5)/(100 - SR + 0.5)) \\ = 6.01 - 0.4756 \times DoF_{positional} - 0.4425 \times DoF_{torsional}$$

Equation 3.4

The equation shows that increasing positional and torsional DoFs has a comparable effect on the reduction of SR.

Recalculating the SR of A34 based on this final ELO model of the positional and torsional DoF, a more accurate prediction of 6% SR is achieved. This prediction is in excellent agreement with the observed 4%SR and demonstrated the better fit of the final ELO model Figure 3.13

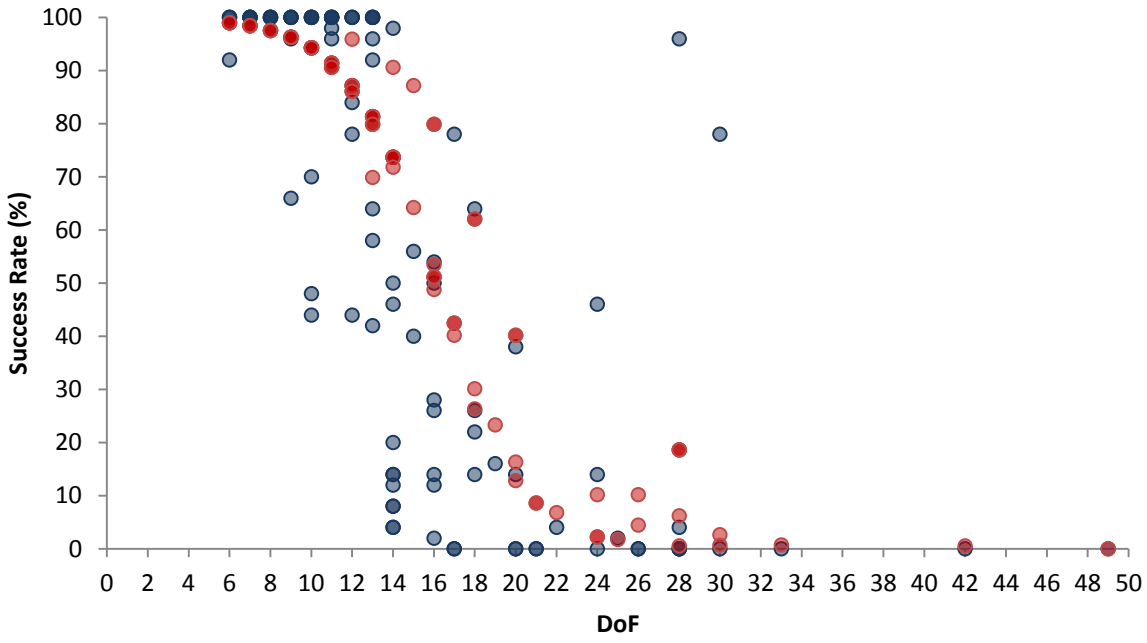


Figure 3.13 The ELO model based on the positional and torsional DoF (red). The observed SRs are given in blue. The depth of shading of each point is proportional to the number structures representative of a particular DoF

Additionally, this final description of the ELO provided the flexibility necessary to describe variation in SR for compounds with the same total DoF, but $DoF_{positional} \neq DoF_{positional}$ and / or

$DoF_{torsional} \neq DoF_{torsional}$. For example, there are 4 compounds with a total of 18 DoF, but different numbers of the positional and torsional DoF (Table 3.9).

Table 3.9 SR information of the crystal structures with 18 DoF.

Data Set No	DoF_{total}	$DoF_{positional}$	$DoF_{torsional}$	$SR_{experimental}$	$SR_{calculated}$
A28	18	12	3	22	26
A29	18	6	6	60	62
A30	18	6	6	34	62
B47	18	6	9	14	30

The variations in the predicted SR values, in the table, clearly demonstrated the difference between the two models, as the ELO based on the total DoF, would predict the same SR for all the entries in Table 3.9.

It may be argued that whilst A28 and A29 show excellent agreement between the experimental and predicted SRs, A30 and B47 do not. Regardless of the large difference between the predicted and experimental SR for B47, the model correctly predicts that the higher torsional DoF of A28 (when compared to A29 and A30) would lead to a significant reduction of the SR, and can be considered an adequate guideline to the required number of SA runs.

The more intriguing observation, however, is the comparison between A29 and A30 as they have the same distribution of DoF and (coincidentally) exhibit the same space group ($P2_1/c$). The data were collected on the same diffractometer and in terms of peak widths, they are very similar. They have both been Pawley fitted to approximately the same resolution and the only obvious difference is the number of reflections fitted, a function of the different unit cell sizes. As such, in theory, it is expected that both crystal structures would generate comparable SR (as predicted by the model). In practice however, their SRs are markedly different and illustrate the difficulties in predicting a-priori success rates for particular problems.

For comparison, an overlay of the calculated SR based on the two models and the experimental baseline results is given in Figure 3.14.

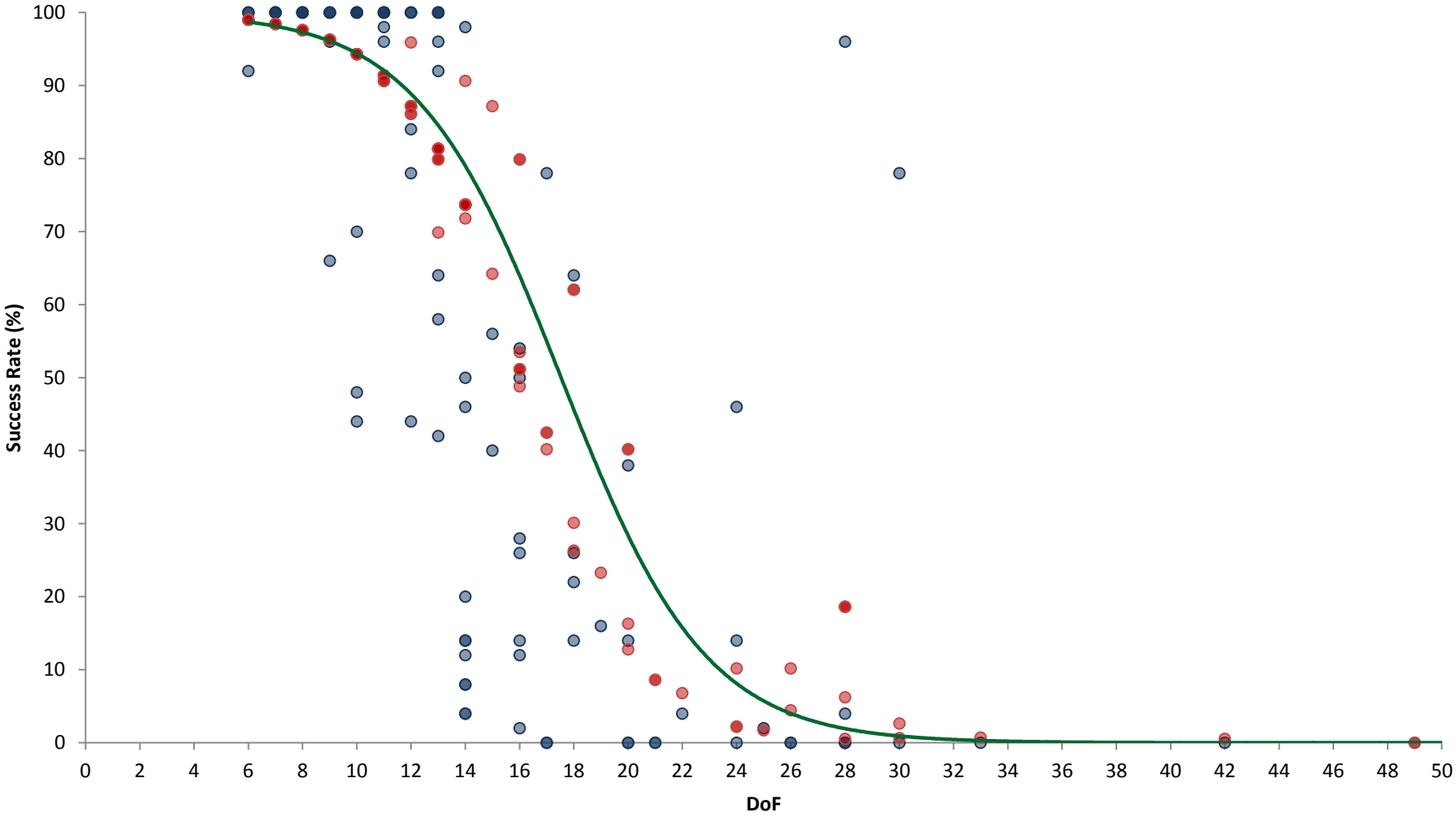


Figure 3.14 An overlay of the observed (blue), and predicted SRs. The SR predicted base on the regression model of the total DoF is given in solid green line, while the predicted SR based on the regression model with both the $DoF_{position}$ and $DoF_{torsional}$ is presented in red.

The limitations of the ELO analysis presented here include that (a) it is based on the FDS, as discussed above; (b) it is based on DASH using 50 SA runs, with 10^7 SA moves and so may not be valid for significantly different numbers of runs / moves. That said, the calculated SR of 0.5% for B60, based on the above model, is in very good agreement with the practically determined SA of 0.4%, even though this SR was based on the 500 SA runs.

Applying the ELO calculation to compound B61 ($DoF_{torsion}$ = 43) predicts a SR of only 0.494×10^{-4}, equating to *ca.* 1 solution in every 20243 SA runs. For this particular structure, this number of runs would need around 2548 days of CPU time (*ca.* 212 days on 12 cores) to execute. Checking the validity of this prediction will therefore need to wait until such time as increased computational resources become available.

Regardless of whether the ELO transform was calculated against the DoF_{total} or against the $DoF_{positional}$ and $DoF_{torsional}$, the predicted SR drop to below 1% for crystal structures with DoF > 30, is consistent with the start of the 'intractable' crystal structure classification used in this work.

3.5 Conclusions

The work described in this chapter represents the most comprehensive testing of an SDPD program carried out to date. EXPO has been tested against a sizeable number of datasets [28 in(Altomare *et al.*, 2008b); 30 in (Altomare *et al.*, 2006); 32 in (Altomare *et al.*, 2008a)], but nothing as comprehensive and focussed on molecular crystal structures as this work. The results give a comprehensive overview of the current DASH capabilities and limitations. 94% of the crystal structures were solved, and only 10% required the use of 500 DASH runs. The fact that only 6 structures remained unsolved by DASH (running on defaults) goes some way to explaining why no effort has been made to date to improve its performance by parameter tuning.

Whilst the ELO was found to give a useful description of the influence of structural complexity upon success rate, there are a number of clear outliers on the SR versus DoF graphs; A7 and B21 for the simple crystal structures, and A38 and B38 for the complex crystal structures. Insights into the reasons for such outliers could possibly be obtained by using the local minimisation approach outlined by Shankland *et al.*, (2010) which allows one to characterise the χ^2 agreement hypersurface in terms of the number of stationary points present and the

frequency with which they are encountered. This approach has the advantage that data resolution and number of reflections are automatically taken into account.

Regardless of the outliers, the following strong recommendations can be made for the future use of DASH, based on the results obtained using the current default SA parameters (*i.e.* T_0 = 0; Cool = 0.02; N_1 = 20; N_2 = 25):

1) For simple crystal structures (DoF < 14), 50 SA runs each performing 5×10^6 SA moves are sufficient to ensure a high level of certainty that the crystal structure has been determined.
2) Moderately complex examples ($14 \geq DoF \leq 20$) require 100 SA runs, each consisting of 10^7 SA moves.
3) 100 SA runs of 10^7 SA moves is also the recommended start point for dealing with complex compounds ($21 \geq DoF \leq 30$). In the event that this does not lead to a solution, either 100 or 500 SA runs (depending upon the available computing power) of 5×10^7 SA moves should be employed
4) 500 SA runs of 5×10^7 SA moves is the absolute minimum for 'intractable' crystal structures with DoF > 30

Finally, it should be noted that the programme of work reported in this chapter alone required a total of 474 days of CPU time.

3.6 References

Altomare A, Caliandro R, Cuocci C, Giacovazzo C, Moliterni AGG, Rizzi R and Platteau C (2008a) Direct methods and simulated annealing: a hybrid approach for powder diffraction data. *J. Appl. Cryst.* **41**:56-61

Altomare A, Cuocci C, Giacovazzo C, Moliterni A and Rizzi R (2008b) Correcting resolution bias in electron density maps of organic molecules derived by direct methods from powder data. *J. Appl. Cryst.* **41**:592-599

Altomare A, Cuocci C, Giacovazzo C, Moliterni AGG and Rizzi R (2006) The combined use of Patterson and Monte Carlo methods for the decomposition of a powder diffraction pattern. *J. Appl. Cryst.* **39**:145-150

Brock CP and Dunitz JD (1994) Towards a grammar of crystal packing. *Chem. Mater.* **6**:1118-1127

Cox DR and Snell EJ (1989) The analysis of bunary data Chapman and Hall, London

Florence AJ, Shankland N, Shankland K, David WIF, Pidcock E, Xu XL, Johnston A, Kennedy AR, Cox PJ, Evans JSO, Steele G, Cosgrove SD and Frampton CS (2005) Solving molecular crystal structures from laboratory X-ray powder diffraction data with DASH: the state of the art and challenges. *J. Appl. Cryst.* **38**:249-259

Kourkoumelis N (2013) PowDLL, a reusable .NET component for interconverting powder diffraction data: Recent developments. *Powder Diffr.* **28**:137-148

Schmidt MU, Ermrich M and Dinnebier RE (2005) Determination of the structure of the violet pigment $C_{22}H_{12}C_{l2}N_6O_4$ from a non-indexed X-ray powder diagram. *Acta Cryst. Sect. B* **61**:37-45

Shankland K, Markvardsen AJ, Rowlatt C, Shankland N and David WIF (2010) A benchmark method for global optimization problems in structure determination from powder diffraction data. *J. Appl. Cryst.* **43**:401-406

Shankland K, McBride L, David WIF, Shankland N and Steele G (2002) Molecular, crystallographic and algorithmic factors in structure determination from powder diffraction data by simulated annealing. *J. Appl. Cryst.* **35**:443-454

Srinivasan R (1991) On the space-group frequency in organic structures. *Acta Cryst. Sect. A* **47**:452

van Langevelde A, Peschar R and Schenk H (2001) Structure of β-trimyristin and β-tristearin from high-resolution X-ray powder diffraction data. *Acta Cryst. Sect. B* **57**:372-377

van Langevelde A, van Malssen K, Driessen R, Goubitz K, Hollander F, Peschar R, Zwart P and Schenk H (2000) Structure of $C_nC_{n+2}C_n$-type (n = even) β'-triacylglycerols. *Acta Cryst. Sect. B* **56**:1103-1111

4 Optimisation of DASH using irace

4.1 Introduction

Following the establishment of the baseline DASH performance, the next objective was the optimisation of the SA control parameters, with the aim of improving the performance of DASH. It was hypothesised that there is a set of SA parameters which will significantly improve the SR and as such lead to the reduction in the ELO coefficients of both the torsional and positional DoF (as given in Chapter 3). Additionally, the optimised parameters would be expected to have a positive impact on the number of steps required to reach a solution, and as such would lead to reductions in the calculation times.

Finding appropriate control parameter values is a challenge for all algorithm (and by extension, software) developers. Whilst the importance of achieving good performance is well recognised, it is not always clear how much effort has been placed into parameter tuning. The SA parameters of DASH, for example, have remained at fixed values since its first release and the performance of DASH, as a function of SA parameter variation, has never been fully investigated. Shankland and co-workers (Shankland *et al.*, 2002) reported some results on DASH performance when varying the initial SA temperature (T_0) and the cooling rate (CR). The results, obtained through testing against a single famotidine data set, showed that the automatic temperature setting routine in DASH was highly effective and that setting the CR value too high (0.3 *c.f.* default value of 0.02) resulted in a drop in SR by a factor of two. However, variation of the parameters N_1 and N_2, which control the way in which SA moves are allocated, was not investigated. Hence, it is conceivable that default SA parameter values in DASH (CR=0.02, N_1=20, N_2=25) are far from optimal, especially when applied to problems of significantly greater complexity than famotidine.

Parameter values may be optimised manually but such manual parameterisation could conceivably lead to poor results due to human bias, unless performed very carefully. A simple example of this (relevant to DASH) would be the bias introduced by the prior work on famotidine and the effect of CR; it is unlikely that any of the researchers involved in that work would go on to explore high values of CR, because of the expectation that it will lead to decreased success rates due to quenching. Thus, despite the fact that their experiments were limited to a single data set and did not vary N_1 and N_2, could conceivably affect the effect of CR upon success rate.

Automatic tuning algorithms ("tuners") on the other hand, can implement the optimisation using approaches which do not require the parameter space to be explored exhaustively and alleviate the problems associated with the human bias in parameter variation.

The design and application of tuners is a dynamic area of research, facilitated by the rapid development of computer technologies. Examples include the work of Eiben and Smith (2012) on tuning evolutionary algorithms, use of SA for the optimisation of mapping on network chips (Yang *et al.*, 2012), mixed integer programming (Hutter *et al.*, 2010), and general-purpose optimisation algorithms (Balaprakash *et al.*, 2007).

The program irace (López-Ibáñez *et al.*, 2011), implementing the iterated racing procedure (as introduced by Balaprakash *et al*.), was used to carry out the SA parameter optimisation of DASH discussed in this chapter. Generally, the main purpose of irace is to find the most appropriate parameter values of an optimisation algorithm (such as the SA algorithm in DASH), given a set of instances typical of that optimisation problem. It has been shown to be most suited for the tuning of metaheuristics (general-purpose optimisation algorithms) of which a characteristic is their relatively large number of configurable parameters of different types, such as the ordered, continuous, categorical and integer. Example can be given with the 'theorem prover' SPEAR (Babić and Hutter, 2008), the ant colony optimisation algorithms applied to the symmetric traveling salesman problem (ACOTS) software package (Stützle, 2002) and the framework of multi-objective ant colony optimisation (MOACO) package (Lopez-Ibanez and Stutzle, 2012), which with their 26, 11 and 16 configurable parameters respectively were used to evaluate the performance of irace (Pérez Cáceres *et al.*, 2014). Additional applications of irace include the configuration of MOACO (López-Ibáñez and Stützle, 2010), the optimisation of an ant colony algorithms in the area of steel production (Fernandez *et al.*, 2015) and the optimisation of state-of-the-art automatic design of robot swarms, achieving the first such method to outperform a human designer (Francesca *et al.*, 2015).

4.1.1 The irace package

Table 4.1 gives definitions for a number of irace related terms used thought this chapter

Table 4.1 Definitions the used irace related terms.

irace term	Symbol	Definition
Parameter space	X	The range of parameter values explored during the optimisation
Tuning instance	i	A representative of the particular optimisable problem (*e.g.* crystal structures)
Training Set	n/a	A set of instances used in irace to benchmark the performance of DASH
Test Set	n/a	A set of instances unseen by irace, used to evaluate the irace results
Configuration	θ_j	A set of the parameter values (*e.g.*CR = 0.02; N_1 = 20; N_2 = 25)
Elite configuration	θ_{elite}	The best performing configuration, output at the end of an iteration
Experiment	n/a	An implementation of the algorithm with a specific configuration
Tuning budget	B	The maximum number of experiments performed

irace is an offline parameter tuner with two clearly defined stages: 1) tuning and 2) evaluation.

The initial tuning stage consists of three phases:

1) Sampling new configurations (*i.e.* sets of parameters) according to a particular distribution (*e.g*. normal distribution or discrete);

2) Selection of the best configurations by means of racing;

3) Updating the sampling distribution in order to bias future iterations towards optimal configurations.

The steps of the tuning stage are repeated until a termination criterion is met - in the case of the current work, this is when the set budget of DASH runs is reached.

The evaluation stage does not involve irace. Rather, it consists of evaluating the performance of each of the elite configurations suggested by irace against a set of instances which were not included in the tuning stage.

A representation of the workflow of SA parameter tuning is given in Figure 4.1. In the figure, the irace 'box' represents the work carried out during the tuning stage. Once all cycles of the tuning are complete, the final elite configurations are output and carried over to the evaluation, which is performed independently of irace.

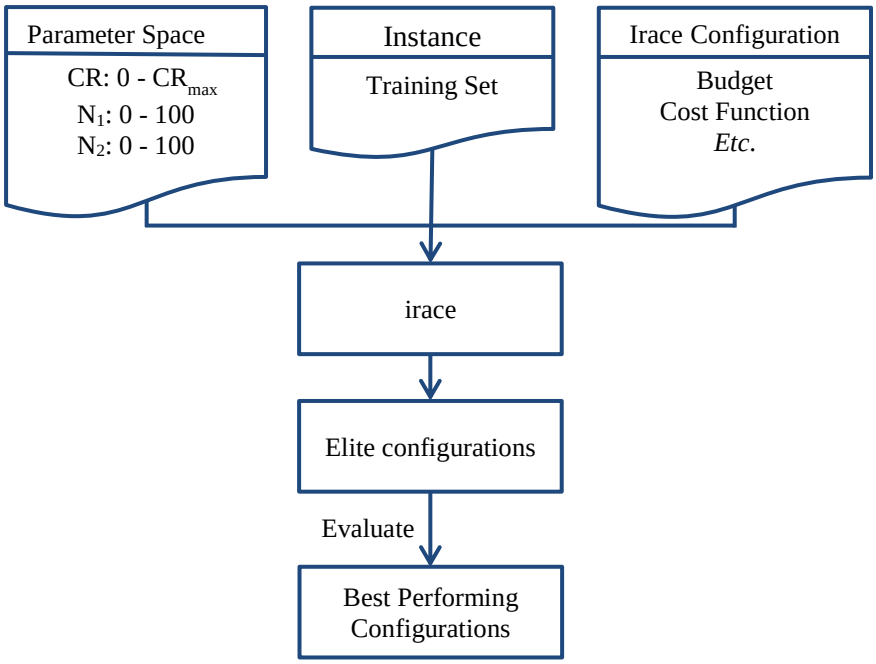


Figure 4.1 The SA parameters tuning workflow

4.1.1.1 The irace input

The required irace inputs as illustrated in Figure 4.1 are: 1) a list of optimisable parameters with their associated parameter space; 2) a set of instances against which the optimisation is performed; and 3) a cost function and additional parameter values.

For the explicit values of the parameter space and the irace configuration, used with the SA parameter optimisation, see section 4.2.1.

4.1.1.2 The iteration

Each repeat of the tuning stage (with its 3 phases) is known as an iteration. The number of iterations performed during an irace run depends upon the number of optimisable parameters, and is calculated using Equation 4.1. Similarly, the budget of each iteration is dependent on the total budget and the number of iterations performed (Equation 4.2).

$$N^{iterations} = 2 + log_2 N^{parameters} \quad \text{Equation 4.1}$$

$$B_j = (B - B_{used})/(N_{iter} - j + 1) \quad \text{Equation 4.2}$$

Once the required inputs are in place, the first iteration starts with the uniform sampling of the parameter space and the generation of a set of parameter configurations [Θ; *e.g*. configuration 1 (CR = 0.20; N_1 = 6; N_2 =11); configuration 2 (CR = 0.22; N_1 = 5; N_2 =25) … configuration n (CR = 0.28; N_1 = 40; N_2 =31)]. Then the race is performed by following the steps given in Table 4.2.

Table 4.2 The steps performed during a single irace iteration.

	irace steps	The DASH equivalent of the irace steps
1.	Evaluate each candidate configuration on the first instance	Perform and evaluate DASH runs against instance 1, looping over the set of parameter configurations (configurations j = 1 to n)
2.	Continue the evaluation on subsequent instances until the number of instances reaches the predefined value of T^{First}	Using the same set of configurations (1 to n), perform and evaluate DASH runs on these (T^{First}-1) instances
3.	Perform statistical test on the evaluated configurations to determine statistically poor performing configurations, if any	Check cost function values to determine which configurations resulted in the poorest DASH performance; for example configuration1
4.	Discard poor performing configurations	Discard configuration 1
5.	Run the next instance with the surviving configurations	Run the next instance with configurations 2 to n
6.	Perform the statistical test every T^{Each} number of instances; T^{Each} is predefined	If T^{Each} = 1, perform statistical test after each instance
7.	Continue until the budget left is insufficient to test all remaining configurations on another instance ($B_j < N_j^{Surviving}$)	Continue until the number of remaining configurations ($N_j^{Surviving}$) is higher than the budget allowed number of DASH runs
8.	Rank the surviving configurations based on their cost function value	Rank the surviving configurations based on their cost function value
9.	Ouput Θ^{Elite} (three by default)	Elite configurations Configuration 6 (CR = 0.16; N_1 = 23; N_2 = 62) Configuration 29 (CR = 0.15; N_1 = 21; N_2 = 46) Configuration 2 (CR = 0.22; N_1 = 5; N_2 = 25)

All subsequent iterations start with the generation of new candidate configurations based upon the elite configurations from the previous iteration.

It is worth noting that the number of candidate configurations generated at the start of an iteration reduces with the increasing number of iterations (Equation 4.3) ensuring more evaluations per configuration are performed in the later iterations.

$$\Theta_j = B_j/(\mu + \min(5, j))$$ Equation 4.3

where μ is a user defined parameter, allowing control over the ratio between the budget and the number of configurations.

When the total budget is exhausted, irace terminates. The top three ranked configurations are then output is an analogous fashion to step 9 in Table 4.2. These are the SA parameter configurations which are then evaluated against the unseen DASH instances (the test set).

4.2 Experimental

4.2.1 The irace calculations

4.2.1.1 Configuring irace

The used configurational settings of irace are summarised in Table 4.3.

Table 4.3. The irace configuration

Iterated racing parameter	irace configuration option	irace/ DASH value
B	maxExperiments	Varied (see Table 4.4)
C (Cost function)	hookRun	$100 \times \chi^2_{profile}/\chi^2_{target}$
μ	mu	5
T^{First}	firstTest	10
T^{Each}	eachTest	1
Statistical test	testType	F-test

4.2.1.2 irace runs

A total of 14 irace calculations were performed, a summary of which is presented in Table 4.4. In all of the runs the optimisable parameters were CR, N_1 and N_2. The CR was varied from 0 to CR_{max} as a real number, while N_1 and N_2 values varied between 0 and 100 as integers. Note that runs with identical irace settings and budgets (*e.g.* runs 1 and 2, 3 and 4 *etc.)* result in different elite configurations, as they have different starting configurations explored, which are tested on different instances, with the use of random SA seed values.

Table 4.4 An outline of the irace runs performed. The results from irace runs can be found in Tables 4.6 and 4.7.

Run Number	Budget	CR_{max}	Molecules used	Subset
1	2500	0.2	A1 - A40	
2	2500	0.2	A1 - A40	
3	5000	0.2	A1 - A40	
4	5000	0.2	A1 - A40	
5	10000	0.2	A1 - A40	
6	10000	0.2	A1 - A40	
7	20000	0.2	A1 - A40	
8	20000	0.2	A1 - A40	
9	30000	0.2	A1 - A40	
10	30000	0.2	A1 - A40	
11	30000	0.3	A18 - A40	Complex
12	30000	0.3	A18 - A40	Complex
13	30000	0.3	A18 - A40	Complex
14	30000	0.3	A18 - A40	Complex

Each irace run consisted of three iterations (in accordance with Equation 4.1), with the exception of runs 12 and 14, where the residual budget from the first three iterations was carried over to perform a fourth iteration.

4.2.2 Evaluation

With 42 elite configurations to address, it was clear from the outset that there were insufficient computational resources to be able to evaluate each configuration against all of the 60 compounds of the test set. Consequently, strategies were adopted to try and reduce the number of required runs to manageable values. Firstly, a custom test set ('the evaluation set', comprising examples from the original test set and some representatives of the training set[12]) was constructed and is shown in Table 4.5. Secondly, attempts were made to identify single configurations that represented a group of configurations *e.g.* 0.25/34/45 and 0.24/32/41 might be considered to be effectively the same and so only one need be evaluated. In practice, such grouping was found to be not successful and instead it was decided that for each irace run, the highest ranked elite configuration should be tested. The only exception was when the highest ranked configuration of an irace run had already been tested from a previous run. In such cases (runs 1 and 8; 2 and 6), the second ranked configuration was evaluated.

[12] Whilst this is not normal practice, it was felt to be important in order to act as 'controls' *i.e.* to cover the possibility that improvements in SR were observed for the training set examples but not for test set examples.

Table 4.5 A summary of the 14 compounds comprising the Evaluation set. The SRs indicated with (*) are a result of the 100 SA runs, whilst the rest are based on 50 SA runs.

No	Compound Name	Total DoF	DoF_{pos}	DoF_{rot}	DoF_{tor}	$SR_{default}$
A20	Famotidine	15	3	3	9	34
A25	Capsaicin	17	3	3	11	2*
A28	Sodium4-[(E)-(4-hydroxyphenyl) diazenyl] benzene sulfonate dihydrate	18	12	3	3	8
A29	Indomethacin:nicotinamide	18	6	6	6	60
A30	Carbamazepine:indomethacin	18	6	6	6	34
A32	S-Ibuprofen	20	6	6	8	18
A34	Verapamil Hydrochloride	22	6	3	13	4
A38	γ-Carbamazepine	28	12	12	4	98
B34	Clarithromycin (I)	14	3	3	8	50
B44	Nimustine hydrochloride	16	6	3	7	8
B47	Tetracaine hydrochloride	18	6	3	9	14
B48	α/β-lactose	20	6	6	8	4*
B52	β-Pigment yellow 191	21	12	3	6	0*
B55	Cytenamide	28	12	12	4	4

Following on from the results of these evaluation experiments, all of the elite configurations of runs 11-14 were evaluated against the evaluation set. (See Table 4.9 for the results)

Finally, 19 semi-arbitrary configurations were constructed from the best performing configurations and evaluated against the evaluation set (see section 4.4.3 for more details on reasoning and Table 4.10 for results).

4.3 Results

4.3.1 irace results

The elite configurations from the 14 irace runs are summarised in Tables 4.6 and 4.7.

Table 4.6 Summary of the elite SA parameters configurations, as a result of the irace calculations, as outlined in Table 4.4 (runs 1-10). (*) indicates the configuration tested.

Run No	Budget (SA runs)	Best SA configurations (CR; N_1; N_2)	CPU time (days)
1	2500	0.19; 44; 13 (*) 0.19; 68; 10 0.19; 49; 15	22
2	2500	0.17; 21; 40 (*) 0.16; 17; 38 0.16; 40; 25	13
3	5000	0.18; 41; 28 (*) 0.2; 22; 39 0.2; 28; 27	34
4	5000	0.2; 26; 29 (*) 0.18; 29; 19 0.18; 64; 8	27
5	10000	0.18; 13; 58 (*) 0.17; 32; 20 0.17; 14; 45	40
6	10000	0.19; 21; 40 0.17; 7; 68 (*) 0.18; 9; 57	45
7	20000	0.16; 18; 38 (*) 0.2; 27; 28 0.19; 26; 28	130
8	20000	0.19; 44; 14 0.18; 11; 64 (*) 0.18; 41; 11	148
9	30000	0.2; 19; 37 (*) 0.19; 30; 19 0.2; 23; 36	172
10	30000	0.2; 34; 21 (*) 0.17; 37; 25 0.2; 40; 23	168

Total CPU time of 799 days.

Table 4.7 Summary of the SA parameters elite configurations, as a result of the irace calculations, as outlined in Table 4.4 (runs 10-14). (*) indicates the configuration tested.

Run No	Budget (SA runs)	Best Sets of SA parameters CR; N_1; N_2	CPU time (days)
11	30000	0.27; 59; 50 (*) 0.25; 31; 56 (*) 0.28; 63; 51 (*)	93
12	30000	0.25; 75; 29 (*) 0.27; 70; 25 (*) 0.26; 74; 23 (*)	115
13	30000	0.25; 46; 62 (*) 0.3; 35; 69 (*) 0.29; 38; 57 (*)	199
14	30000	0.24; 18; 84 (*) 0.21; 16; 85 (*) 0.21; 19; 91 (*)	109

Total CPU time of 516 days.

4.3.2 Configuration evaluation

The results from the evaluation experiments as outlined in section 4.2.2 are presented here.

Table 4.8 The evaluation of the highest ranked elite configurations of the irace runs 1-10. The SRs achieved with the default SA parameters are marked with green, whilst the best performing configuration is highlighted in light blue. * indicates the reported SR is a result of the 100 SA runs.

	SR (%)													
SA parameters CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.19; 44; 13	24	8	22	44	12	16	0	58	62	2	10	4	0	60
0.17; 21; 40	32	6	26	58	22	22	10	66	72	4	30	8	2	60
0.18; 41; 28	54	14	20	76	30	30	14	62	80	10	18	2	0	58
0.20; 26; 29	34	2	28	40	16	22	10	54	50	6	24	14	0	46
0.18; 13; 58	28	2	14	18	12	20	8	56	50	4	8	4	2	68
0.17; 7; 68	20	4	16	30	10	34	10	36	38	6	30	2	2	38
0.16; 18; 38	28	6	14	42	12	28	10	58	52	6	20	6	0	28
0.18; 11; 64	36	4	14	52	8	0	6	64	60	4	2	10	0	36
0.20; 19; 37	34	2	24	40	14	0	10	44	56	10	2	6	2	38
0.20; 34; 21	26	6	6	0	46	2	4	58	42	4	20	8	2	44

Table 4.9 The evaluation of the elite configurations of the irace runs 11-14. The SRs achieved with the default SA parameters are marked with green, whilst the best performing configuration is highlighted in light blue. * indicates the reported SR is a result of the 100 SA runs.

	SR (%)													
SA parameters CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.27; 59; 50	66	14	24	80	26	0	18	90	94	30	52	6	6	78
0.25; 31; 56	58	8	18	78	34	34	14	90	88	12	52	2	8	66
0.28; 63; 51	84	14	20	84	46	0	40	84	98	30	58	10	12	96
0.25; 75; 29	62	26	26	80	36	36	22	96	92	18	48	6	4	82
0.27; 70; 25	52	6	28	60	14	38	12	84	88	12	36	14	0	88
0.26; 74; 23	44	2	24	72	22	40	10	16	90	14	52	14	4	84
0.25; 46; 62	72	26	30	82	26	48	24	98	96	26	62	8	16	90
0.30; 35; 69	48	10	32	76	36	48	20	20	88	16	62	12	4	54
0.29; 38; 57	68	8	24	64	28	0	24	70	90	26	64	10	10	86
0.24; 18; 84	52	2	20	66	20	28	12	56	82	10	44	16	6	74
0.21; 16; 85	44	8	20	66	72	32	20	60	86	14	50	4	10	70
0.21; 19; 91	58	10	32	82	86	32	20	16	96	22	50	6	0	86

Table 4.10 The evaluation of the additional, semi-arbitrary test configurations. † indicates manual T_0=13(T_0 = 0 caused failure with A38), see section 4.4.3.3 for more details. The SRs achieved with the default SA parameters are marked with green, whilst the best performing configuration is highlighted in light blue. * indicates the reported SR is a result of the 100 SA runs.

	SR(%)													
SA parameters CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.27; 73; 56	88	24	40	96	56	54	36	100†	100	48	54	12	18	90
0.27; 63; 51	78	14	26	88	38	32	20	96†	100	16	68	10	14	98
0.27; 60; 63	66	10	32	94	41	48	26	100†	96	36	64	4	8	96
0.27; 73; 61	92	12	44	90	50	42	34	100†	100	52	64	8	18	64
0.27; 73; 51	90	18	32	92	36	56	44	100†	98	26	58	14	10	96
0.27; 73; 41	72	12	40	92	44	44	30	96†	96	32	64	8	8	90
0.27; 59; 63	74	16	24	96	38	56	22	98†	100	28	56	8	12	96
0.27; 53; 61	54	10	26	74	22	62	26	98†	100	34	64	8	6	0
0.27; 53; 51	64	10	44	86	28	42	22	96†	96	32	38	4	10	90
0.27; 49; 40	62	10	26	76	22	41	22	80†	94	18	50	10	10	82
0.27; 20; 73	44	6	28	66	16	26	12	58	84	6	40	10	10	70
0.27; 73; 20	50	6	20	64	24	28	20	48	88	14	54	10	8	76
0.25;31; 66	58	6	28	82	26	44	24	90†	100	14	58	14	4	98
0.25; 31; 76	78	18	20	84	28	44	22	96†	92	24	26	74	6	90
0.25; 31; 86	68	16	20	86	40	26	18	96†	98	20	54	8	8	90
0.27; 35; 86	82	16	44	94	32	46	22	98†	98	32	62	4	12	96
0.19; 20; 73	62	4	22	86	40	34	22	82	90	10	38	10	4	78
0.19; 73; 20	54	4	34	74	16	38	10	82	90	14	44	8	6	2
0.19; 25; 63	50	6	20	64	24	28	20	48	88	14	54	10	8	76

4.4 Discussion

A number of important considerations were made in the initial steps of the irace setup, in order to prepare the required inputs:

1) **Training set homogeneity**: The division of the FDS into the training and test sets was originally implemented to satisfy irace's need for two sets of instances – one which is used during the parameter optimisation and the other for the purpose of evaluating irace's suggested elite configurations.

   The 40 crystal structures of the training set (A1-A40) (see Chapter 2) were selected as being representative of crystal structures in the FDS and more broadly, of problems likely to be tackled with DASH. As such, this training set consisted of compounds of all the complexity groups (DoF varied between 6 and 40).

   Here, the question of the set's homogeneity emerged. The homogeneity of the training instances has previously been recognised as an important factor governing the quality of the tuning outcome (Pérez *et al.*, 2013; Schneider and Hoos, 2012). To paraphrase the results of these studies, it may not be possible to produce a single set of optimised parameters that performs well with the full range of complexities presented in the training set. In order to account for this possibility, whilst initial irace runs (run 1-10) were carried out on the full heterogeneous training set, the subsequent irace runs (runs 11-14) were carried out on a subset of 'complex structures' excised from the full training set.

2) **Which SA parameters should be optimised**? DASH has very few parameters which determine its performance. These are the starting temperature (T_0), the cooling rate (CR), and N_1 and N_2, the product of which governs the number of steps performed at each temperature, before the cooling step is applied.

   Currently a value of '0' is used for T_0, which tells DASH to automatically determine an optimal value of this parameter for the specific example at hand. This is achieved by performing a short preliminary SA run during which the deviation of χ^2 at different temperatures is explored. The temperature above which no significant variations in the χ^2 values are observed is selected as initial. Whilst manual setting of T_0 is feasible, a low initial value is likely to prevent the SA escaping local minima and a high T_0 value will simply result in wasted SA moves. Given that automatic setting of T_0 is done on a sound

scientific basis, it was decided not to include T_0 as an optimisable parameter in the irace calculations. The remaining parameters, CR, N_1 and N_2 were selected for optimisation.

Here an important consideration was the size of the explored parameter space. Large parameter ranges require a large irace budget to ensure completeness, *i.e.* a thorough investigation is carried out; however conservative ranges may result in missed opportunities. In this work, the parameter ranges were set pragmatically, recognising that they needed to accommodate significant changes from the default DASH parameters but also acknowledging the computational requirements of spanning large areas of parameter space (specific values Section 4.2.1.2).

3) **The cost function**. This was defined as in Equation 4.4 given below, where the target χ^2 value, was representative of the solved crystal structure.

$$100 \times \chi^2_{profile} / \chi^2_{target} \qquad \text{Equation 4.4}$$

In order to correctly establish the target value, a rigid-body Rietveld refinement of the previously deposited crystal structure was performed with DASH. The χ^2 value of the refinement was assumed to be the lowest achievable during the SA and as such was chosen as the target.

4) **The budget**. There was some uncertainty as to whether irace runs of large numbers of DASH calculations would give superior results to their small budget counterparts. Calculations with large budgets (*e.g.* 30,000 DASH runs) were generally expected to give better results, due to the larger number of evaluations carried out. Ultimately, to take account of the stochastic nature of irace, and to explore all options, irace runs of varying budgets were performed (see Table 4.4).

Table 4.11 presents a summary of the steps carried out during the irace optimisation and the rationale behind them. Further details of each step are discussed below.

Table 4.11 A summary of the work carried out in this chapter. Unless otherwise stated, all evaluation steps were performed over the Evaluation set summarised in Table 4.5.

Step	Experiment	Rationale	Outcome	Outcome
1	Perform 10 irace runs (runs 1-10,Table 4.4)	Seek optimal SA configuration/s	10 sets of 3 elite configurations of the SA parameters	Table 4.6
2	Evaluate the configurations resulting from step 1, by grouping them	Fast evaluation of the quality of the elite configuration of step 1	No significant SR improvements. The grouping evaluation method was recognised to be inappropriate.	Not shown
3	Evaluate the configurations resulting from step 1, by testing the highest ranked elite configurations of each irace run (1-10)	Fast evaluation of the quality of the elite configuration of step 1	No significant improvements in the SR is observed	Table 4.8
4	Perform further 4 irace runs (runs 11-14, Table 4.4)	Explore new SA parameter configurations	4 sets of 3 elite configurations of the SA parameters	Table 4.7
5	Evaluate the configurations of step 4, by testing the best ranked elite configurations of each irace run (11-14)	Fast evaluation of the quality of the new elite configuration	Good overall improvements in the SR observed.	Table 4.9
6	Evaluate the remaining elite configurations of irace runs 11-14	Explore all elite configurations of irace runs 11-14	Excellent improvements with specific configurations; exception A32	Table 4.9
7	Derive new configurations, based on best performing so far	Explore the performance of A32 with new, potentially well performing, configurations	New SA parameter configurations	Table 4.15
8	Evaluate the configurations derived in step 7 over A32 only	Find configurations well performing with A32	8 configurations, performing well with A32	Table 4.15
9	Evaluate the best configurations of step 8 over the Evaluation set	Evaluate the quality of the new configurations	Excellent improvements with specific configurations	Table 4.10 Table 4.16
10	Establish the best performing configurations	Determine the optimal performing configurations	Six well performing configurations	Table 4.17

4.4.1 Runs 1-10

The results from the irace runs 1-10 (Table 4.6) showed a CR typically at the higher end of the allowed maximum bound (CR_{max} = 0.2), and an average value of 0.19. Based on previous DASH successes with the default cooling rate of 0.02, such high CR values were somewhat surprising. Indeed, the expectation was that high values of CR would lead to 'quenching' during the SA, which was demonstrated by the simple test shown in Table 4.12. Nevertheless, the consistently high CR values in all elite configurations returned by runs 1-10 strongly suggested that the upper bound, CR_{max}, needed to be extended.

Table 4.12 Assessment of a test configuration devised from the combinations of aggressive and default SA parameter values. The baseline SR is shown in green. * indicates the reported SR is a result of the 100 SA runs.

	SR (%)													
CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.17; 20; 25	6	0	0	4	0	12	0	0	2	0	2	2	0	0

Irrespective of budget, individual values of N_1 and N_2 varied markedly with no obvious trend. Initially, some attempt was made to group 'similar' elite configurations, in order to reduce the number of configurations that needed to be fully evaluated. The product $N_1 \times N_2$ was used to group the configurations, as it is a direct determinant of the number of SA step performed at each temperature level. However, the evaluation of the six group representative configurations created showed poor results throughout, with little or no increase in the observed SR across the evaluation set. Following the advice of Manuel López-Ibáñez, and returning to the original idea of avoiding human bias, the configuration evaluation process proceeded using only the exact configurations suggested by irace. In order to manage computational requirements, only the first configuration (*i.e.* the highest ranked) from each irace run was initially tested, unless otherwise indicated.

From the 10 tested SA parameter configurations, the best overall improvement was given by the configuration 0.18/41/28 (highlighted in blue, Table 4.8). However, it was notable that whilst some compounds (*e.g.* A20, A28, A34, and B55) saw increased SR, others (A30, A38 and B47) saw decreases and B52 remained unsolved. This indicated that this best configuration was far from optimal and that further irace calculations were needed. The possibility that the training set displayed excessive heterogeneity was considered to be a likely factor contributing to the relative failure of runs 1-10.

4.4.2 Runs 11-14

To address the limitations exposed by runs 1-10, runs 11-14 utilised $CR_{max} = 0.3$; and a training set with greater homogeneity *i.e.* only the compounds (A18-A40) with DoF > 13 were used.

The assessment of the best-ranked elite configurations against the evaluation set presented a much more encouraging set of results (see Table 4.13). Significantly, the CR was greater than 0.2 for all the elite configurations but comfortably less than the 0.3 upper limit. The 0.25/46/62 configuration stands out, increasing the SR by factor of at least two for 10 out of the 14 compounds. A38 retained its high SR of 98%, and only A30 displayed a small reduction in SR

from 34% to 26%. Given the success of the best-ranked configurations, the remaining elite configurations for irace runs 11-14 were also evaluated (Table 4.14)

Table 4.13 Evaluation results of the first elite configurations of irace runs 11-14. The highest achieved SRs are highlighted in dark blue, the second highest in light blue and the rest of the improved SRs are marked in grey. * indicates the reported SR is a result of the 100 SA runs.

	SR (%)													
CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.27; 59; 50	66	14	24	80	26	0	18	90	94	30	52	6	6	78
0.25; 75; 29	62	26	26	80	36	36	22	96	92	18	48	6	4	82
0.25; 46; 62	72	26	30	82	26	48	24	98	96	26	62	8	16	90
0.24; 18; 84	52	2	20	66	20	28	12	56	82	10	44	16	6	74

Table 4.14 Evaluation results of all elite configurations of irace runs 11-14. The highest achieved SRs are highlighted in dark blue, the second highest in light blue and the rest of the improved SR are marked in grey. Please note that the results presented here are identical to those of Table 4.9, but with a different colour coding.

	SR (%)													
CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4
0.27; 59; 50	66	14	24	80	26	0	18	90	94	30	52	6	6	78
0.25; 31; 56	58	8	18	78	34	34	14	90	88	12	52	2	8	66
0.28; 63; 51	84	14	20	84	46	0	40	84	98	30	58	10	12	96
0.25; 75; 29	62	26	26	80	36	36	22	96	92	18	48	6	4	82
0.27; 70; 25	52	6	28	60	14	38	12	84	88	12	36	14	0	88
0.26; 74; 23	44	2	24	72	22	40	10	16	90	14	52	14	4	84
0.25; 46; 62	72	26	30	82	26	48	24	98	96	26	62	8	16	90
0.30; 35; 69	48	10	32	76	36	48	20	20	88	16	62	12	4	54
0.29; 38; 57	68	8	24	64	28	0	24	70	90	26	64	10	10	86
0.24; 18; 84	52	2	20	66	20	28	12	56	82	10	44	16	6	74
0.21; 16; 85	44	8	20	66	72	32	20	60	86	14	50	4	10	70
0.21; 19; 91	58	10	32	82	86	32	20	16	96	22	50	6	0	86

Two configurations stand out: one is the previous best performing configuration of 0.25/46/62 and the other is 0.28/63/51, which was the third ranked configuration of run 11. The former has the best overall performance, whilst the latter displays the largest number of individual gains but also a conspicuous failure - compound A32. Attempts were then made to retain the high SRs of 0.28/63/51 whilst eliminating the problem of A32.

4.4.3 Additional calculations

4.4.3.1 Compound A32 - Ibuprofen

A series of semi–arbitrary test configurations were established based on combinations of the CR, N_1 and N_2 values from the elite configurations of runs 11-14. These were first tested on A32 alone and configurations with SR above 40% were then assessed over the evaluation set (Table 4.15).

Table 4.15 Evaluation of additional test configurations.

	Test SA parameter configurations												
	0.27 73;56	0.27 73;61	0.27; 73;41	0.27 63;51	0.27 65;59	0.27; 63;61	0.27 79;14	0.27 56;54	0.27 53;61	0.27 50;59	0.27 59;63	0.27 60;63	0.27 73;51
SR (%)	54	42	44	32	0	0	24	38	62	2	56	48	56

Notably, use of 0.27/63/51 (a seemingly trivial change from 0.28/63/51) resulted in SR increase from 0% to 32%. Whilst a positive change, it may be indicative of the unpredictability of SR change as a function of SA parameters. Although this configuration did not quite reach the pre-set criterion of 40% SR, it was nevertheless carried forward as one of the eight configurations selected for further assessment over the whole evaluation set.

4.4.3.2 Evaluation of the additional runs

In addition to the eight configurations derived based on the Ibuprofen testing, a further 11 new SA parameter configurations were devised to promote the exploration of the previously observed values of CR, N_1 and N_2. Their generation was performed in a random manner, and had the sole purpose of exploring for a trend of how the SR changes as a result of the variation of the three SA parameters. The assessment of these 19 configurations against the Evaluation set is summarised in Table 4.10, whilst the colour coded results and some statistical values, are presented in Table 4.16.

Table 4.16 Evaluation results of the additional test configurations. The highest achieved SRs are highlighted in dark blue, whilst the second highest are in light blue, and the rest of the improved SR are marked in grey. † indicates manual T_0=13(T_0 = 0 caused failure with A38). 'Average All' = the SR average of the 14 evaluation compounds; 'Average A' = the SR average of the training set compounds (A20- A38); 'Average B' = the SR average of the test set compounds (B34- B55). The selected 5 configurations are highlighted in orange.

	SR (%)														Statistical Analysis			
CR; N_1; N_2	A20	A25	A28	A29	A30	A32	A34	A38	B34	B44	B47	B48	B52	B55	Average All	Average A	Average B	Median
0.02; 20; 25	34	2*	8	60	34	18	4	98	50	8	14	4*	0*	4	24.1 ±28.5	32.3 ±39.9	13.3 ±18.6	11
0.27; 73; 56	88	24	40	96	56	54	36	100†	100	48	54	12	18	90	58.3 ±31.3	61.8 ±29.2	53.7 ±36.1	54
0.27; 63; 51	78	14	26	88	38	32	20	96†	100	16	68	10	14	98	49.9 ±31.3	49.0 ±33.2	51.0 ±42.8	35
0.27; 60; 63	66	10	32	94	41	48	26	100†	96	36	64	4	8	96	51.5 ±34.8	52.1 ±32.1	50.7 ±41.2	44.5
0.27; 73; 61	92	12	44	90	50	42	34	100†	100	52	64	8	18	64	55.0 ±31.6	58.0 ±27.8	51.0 ±33.7	51
0.27; 73; 51	90	18	32	92	36	56	44	100†	98	26	58	14	10	96	55.0 ±34.1	58.5 ±29.2	50.3 ±39.9	50
0.27; 73; 41	72	12	40	92	44	44	30	96†	96	32	64	8	8	90	52.0 ±32.9	53.8 ±28.6	49.7 ±39.4	44
0.27; 59; 63	74	16	24	96	38	56	22	98†	100	28	56	8	12	96	51.7 ±35.2	53.0 ±34.3	50.0 ±40.9	47
0.27; 53; 61	54	10	26	74	22	62	26	98†	100	34	64	8	6	0	41.7 ±33.7	46.5 ±31.5	35.3 ±39.7	30
0.27; 53; 51	64	10	44	86	28	42	22	96†	96	32	38	4	10	90	47.3 ±33.3	49.0 ±30.7	45.0 ±39.4	40
0.27; 49; 40	62	10	26	76	22	41	22	80†	94	18	50	10	10	82	43.1 ±30.5	42.4 ±26.9	44.0 ±37.3	33.5
0.27; 20; 73	44	6	28	66	16	26	12	58	84	6	40	10	10	70	34.0 ±26.5	32.0 ±22.4	36.7 ±33.8	27
0.27; 73; 20	50	6	20	64	24	28	20	48	88	14	54	10	8	76	36.4 ±26.8	32.5 ±18.0	41.7 ±35.7	26
0.25; 31; 66	58	6	28	82	26	44	24	90†	100	14	58	14	4	98	46.1 ±34.8	44.8 ±29.6	48.0 ±43.7	36
0.25; 31; 76	78	18	20	84	28	44	22	96†	92	24	26	74	6	90	50.1 ±33.3	48.8 ±33.1	52.0 ±37.7	36
0.25; 31; 86	68	16	20	86	40	26	18	96†	98	20	54	8	8	90	46.3 ±34.7	46.3 ±34.6	46.3 ±40.7	33
0.25; 35; 86	82	16	44	94	32	46	22	98†	98	32	62	4	12	96	52.7 ±35.0	54.3 ±32.2	50.7 ±41.1	45
0.19; 20; 73	62	4	22	86	40	34	22	82	90	10	38	10	4	78	41.6 ±32.0	44.0 ±29.0	38.3 ±37.5	36
0.19; 73; 20	54	4	34	74	16	38	10	82	90	14	44	8	6	2	34.0 ±30.7	39.0 ±29.7	27.3 ±34.2	25
0.19; 25; 63	50	6	20	64	24	28	20	48	88	14	54	10	8	76	36.4 ±26.8	32.5 ±18.0	41.7 ±35.7	26

The highlighting of the highest SRs identifies the 5 best-performing configurations: 0.27/73/56, 0.27/60/63, 0.27/73/51, 0.27/59/63 and 0.25/35/85. As expected for so few compounds possessing such wide structural diversity, the average SRs possessed very high standard deviations and did not provide additional insights beyond those obtained "by eye". Configuration 0.29/59/63 was not included as part of the best-performing set, although its exclusion was marginal.

As such the 5 best-performing SA parameter configurations, together with configuration 0.25/46/62 (identified in Table 4.10 as the best performing in runs 11-14) were selected as the best performing *overall* and their performance is summarised in Table 4.17. It is these six configurations that were then tested against the FDS (see Chapter 5).

Table 4.17 Summary of the best performing SA parameter configurations and their average SR. Average All = the average based on the SR of all 14 evaluation compounds; average A = the average based on the SRs of the training set representative compounds (A20-A38); and average B = the average based on the SRs of the test set representative compounds (B34-B55). †The reported SR was achieved with a manually set T_0 value of 13. * indicates the default SR achieved with 100 SA runs. ‡ The average values are reported with no standard deviations due to their irrelevance (as discussed above).

Compound	SR (%)						
	0.02; 20; 25	0.27; 73; 56	0.27; 60; 63	0.27; 73; 61	0.27; 73; 51	0.25; 46; 62	0.25; 35; 86
A20	34	88	66	92	90	72	82
A25	2*	24	10	12	18	26	16
A28	8	40	32	44	32	30	44
A29	60	96	94	90	92	82	94
A30	34	56	41	51	36	26	32
A32	18	54	48	42	56	48	46
A34	4	36	26	34	44	24	22
A38	98	100†	100†	100	100†	98	98†
B34	50	100	96	100	98	96	98
B44	8	48	36	52	26	26	32
B47	14	54	64	64	58	62	62
B48	4*	12	4	8	14	8	4
B52	0*	18	8	18	10	16	12
B55	4	90	96	64	96	90	96
Average‡ All	24.1	58.3	51.5	55.1	55.0	50.3	52.7
Average‡ A	32.3	61.8	52.1	58.1	58.5	50.8	54.3
Average‡ B	13.3	53.7	50.7	51.0	50.3	49.7	50.7
Median	11	54	44.5	51	50	39	45

4.4.3.3 The γ-carbamazepine exception

The evaluation of a number of test configurations (derived from irace and/or further manually generated) against A38 (γ-carbamazepine) resulted in a success rate of zero. Originally, this was attributed to poor performance of the chosen SA parameters. On closer inspection

however, it was noticed that the DASH runs were terminating with very short calculation times. For example, the use of both 0.27/60/63 and 0.27/73/56 gave DASH runs that completed in just over 2 hours, while the baseline DASH calculations took 17 hours. This abnormality prompted further investigations of the problematic configurations and the issue was finally traced to an array overflow error in the SA code. This overflow caused T_0 to be set (literally) to a value of zero, leading to immediate termination of the SA algorithm (*i.e.* no SA moves performed) and invocation of the terminal simplex, which was solely responsible for the DASH output χ^2 value.

For these cases, a manually set value, $T_0 = 13$, was used instead of the default. SR values achieved in such manner are denoted by † in the relevant tables. The T_0 value of 13 was selected as appropriate for γ-carbamazepine based on observations from the baseline SA runs.

Since the discovery of the γ-carbamazepine exception, this error has been found with a small number of other compounds which do not fall into the scope of this thesis. From these examples, it was concluded that the issue is related to the high number of atoms in the asymmetric unit, and the crystal symmetry. The SA code has now been patched by CCDC, though a more detailed fix remains to be made. The new SA code was tested to ensure it gives results comparable with the $T_0 = 13$ DASH runs (see Table 4.18)

Table 4.18 Comparison between the results the current DASH algorithm and the adjusted DASH algorithm for the purpose of A38.

	$DASH_{original}$ SR (%)		$DASH_{adjusted}$ SR (%)	
	0.27; 73; 56	0.27; 60; 63	0.27; 73; 56	0.27; 60; 63
A34 (Verapamil)	36	26	34	36
A38 (γ-carbamazepine)	100†	100†	100	98

† SR achieved with T_0 manually set to 13

4.5 Conclusions

The irace runs have returned a set of six SA parameter configurations which show significant gains in performance (with respect to DASH default SA settings) over the fourteen compounds of the evaluation set. These configurations are henceforth referred to as the "aggressive" SA parameters sets, or the "aggressive" configurations. The general applicability of these settings to the full data set is addressed in Chapter 5. It seems unlikely that any of the selected configurations would have been 'chosen' without the use of irace, particularly in light of the high CR which features in all six configurations.

Multiple linear regression was also employed to try and establish a correlation between the SR of a specific compound and the values of the SA parameters, but none could be found.

Inevitably the work in this chapter has involved some compromises in experimental design in order to limit computational requirements. Despite this, the irace calculations *alone* utilised 225,000 DASH run requiring approximately 1315 CPU days. Thus whilst it is quite possible that better performing configurations may have been identified by more extensive/additional irace runs, it can be confidently concluded that the results obtained here are the best attainable given the available computing power and time.

4.6 References

Babić D and Hutter F (2008) Spear theorem prover, in *SAT'08: Proceedings of the SAT 2008 Race*

Balaprakash P, Birattari M and Stützle T (2007) Improvement strategies for the F-Race algorithm: Sampling design and iterative refinement, in *Hybrid Metaheuristics* (Bartz-Beielstein T, Blesa Aguilera M, Blum C, Naujoks B, Roli A, Rudolph G and Sampels M eds) pp 108-122, Springer Berlin Heidelberg

Eiben AE and Smit SK (2012) Evolutionary Algorithm Parameters and Methods to Tune Them, in *Autonomous Search* (Hamadi Y, Monfroy E and Saubion F eds) pp 15-36, Springer Berlin Heidelberg

Fernandez S, Alvarez S, Malatsetxebarria E, Valledor P and Daz D (2015) Performance comparison of ant colony algorithms for the scheduling of steel production lines, in *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation* pp 1387-1388, ACM, Madrid, Spain

Francesca G, Brambilla M, Brutschy A, Garattoni L, Miletitch R, Podevijn G, Reina A, Soleymani T, Salvaro M, Pinciroli C, Mascia F, Trianni V and Birattari M (2015) AutoMoDe-Chocolate: automatic design of control software for robot swarms. *Swarm Intell.* **9**:125-152

Hutter F, Hoos H and Leyton-Brown K (2010) Tradeoffs in the empirical evaluation of competing algorithm designs. *Ann. Math. Artif. Intell.* **60**:65-89

López-Ibáñez M, Dubois-Lacoste J, Stützle T and Birattari M (2011) The irace package, Iterated Race for Automatic Algorithm Configuration. Université libre de Bruxelles, IRIDIA, TR/IRIDIA/2011-004, Belgium, http://iridia.ulb.ac.be/irace/, Oct 2015

Lopez-Ibanez M and Stutzle T (2012) The Automatic Design of Multiobjective Ant Colony Optimization Algorithms. *Evolutionary Computation, IEEE Transactions on* **16**:861-875

López-Ibáñez M and Stützle T (2010) Automatic Configuration of Multi-Objective ACO Algorithms, in *Swarm Intell.* (Dorigo M, Birattari M, Di Caro G, Doursat R, Engelbrecht A, Floreano D, Gambardella L, Groß R, Şahin E, Sayama H and Stützle T eds) pp 95-106, Springer Berlin Heidelberg

Pérez Cáceres L, López-Ibáñez M and Stützle T (2014) An Analysis of Parameters of irace, in *Evolutionary Computation in Combinatorial Optimisation* (Blum C and Ochoa G eds) pp 37-48, Springer Berlin Heidelberg

Pérez L, López-Ibáñez M and Stützle T (2013) An Analysis of Parameters of irace. Université libre de Bruxelles, Series I-TR, TR/IRIDIA/2013-014,

Schneider M and Hoos H (2012) Quantifying Homogeneity of Instance Sets for Algorithm Configuration, in *Learning and Intelligent Optimization* (Hamadi Y and Schoenauer M eds) pp 190-204, Springer Berlin Heidelberg

Shankland K, McBride L, David WIF, Shankland N and Steele G (2002) Molecular, crystallographic and algorithmic factors in structure determination from powder diffraction data by simulated annealing. *J. Appl. Cryst.* **35**:443-454

Stützle T (2002) ACOTSP: A software package of various ant colony optimization algorithms applied to the symmetric traveling salesman problem, http://www.aco-metaheuristic.org/aco-code/, Oct 2015

Yang B, Guang L, Säntti T and Plosila J (2012) Parameter-optimized simulated annealing for application mapping on Networks-on-Chip, in *Learning and Intelligent Optimization* (Hamadi Y and Schoenauer M eds) pp 307-322, Springer Berlin Heidelberg

5 Optimised DASH – implementation of aggressive parameter settings

5.1 Introduction

This chapter evaluates the performance of DASH with the six aggressive SA parameter configurations determined using the irace runs described in Chapter 5. Different performance characteristics were considered, with the success rate of crystal structure solution being the primary selection criterion. The CPU time required to reach a solution and the accuracy of the final crystal structure solutions was also considered, with the overall aim of identifying the best performing, general purpose, configuration for future implementation in DASH.

5.2 Experimental

The six aggressive SA parameter configurations ($CR/N_1/N_2$ = 0.27/73/56, 0.27/73/61, 0.27/73/51, 0.27/60/63, 0.25/35/86 and 0.25/46/62) were tested against all molecules in the FDS. The DASH runs performed using each configuration mirrored the DASH baseline calculations, *i.e.* initially 50 SA runs of 10^7 steps were performed for all molecules, followed by 100 SA runs of 10^7 steps for the unsuccessful examples. Finally 500 SA runs of 5×10^7 steps were carried out for the required compounds.

To further facilitate the direct comparison of the results, all DASH runs were performed in identical manner to those of the baseline, *i.e*. identical molecular models were used for the generation of z-matrices, all torsion angles were freely rotatable (from 0^o to 360^o) during the SA calculations, the same random seeds values were used and a value of one was selected for the χ^2 multiplier to ensure all SA moves are executed. As such, any observed changes in SR are a consequence of the SA parameter configuration[13].

To accelerate the identification of successful solutions, the highest χ^2 value deemed to be successful for the baseline runs was used as a guideline. For example, a χ^2 upper limit of 35 for compound A34 meant that when evaluating the results of the aggressive parameter runs, all DASH solutions of χ^2 up to 35 were automatically accepted as successful; any solutions with χ^2 slightly greater than 35 were examined against the reference crystal structure.

5.3 Results

Table 5.1 and Table 5.2 summarise the results achieved with the six aggressive configurations. The colour coding of the SR results used in Chapter 4 was found to be particularly helpful in

[13] Bearing in mind the stochastic nature of the SA.

identifying the best performing configurations, and as such was also applied to the results reported here.

Table 5.1 A summary of the SRs achieved with the six aggressive SA parameter configurations, against the FDS, based on the 50 and 100 SA runs. * indicates the SR achieved with 100 SA runs. † indicates SRs achieved with T_0 manually set to 13 (as discussed in Section 4.4.3.2).

No	Default DASH	0.27; 73;56	0.27; 73;61	0.27; 73;51	0.27; 60; 63	0.25; 35; 86	0.25; 46; 62
A1	100	100	100	100	100	100	100
A2	100	100	100	100	100	100	100
A3	100	100	100	100	100	100	100
A4	100	100	100	100	100	100	100
A5	100	100	100	100	100	100	100
A6	100	100	100	100	100	100	100
A7	48	78	78	70	62	58	74
A8	100	100	100	100	100	100	100
A9	100	100	100	100	100	100	98
A10	100	100	100	100	100	100	100
A11	100	100	100	100	100	100	100
A12	100	100	100	100	100	100	100
A13	78	98	98	96	94	94	92
A14	96	100	100	100	100	100	100
A15	100	100	100	100	100	100	100
A16	42	74	96	82	86	82	90
A17	100	100	100	100	100	100	100
A18	4	6	8	4	2	6	2
A19	14	12	10	14	14	12	24
A20	34	88	92	90	66	82	72
A21	56	78	98	92	86	74	78
A22	28	74	86	70	62	64	52
A23	54	92	88	70	72	82	76
A24	50	84	92	78	80	86	80
A25	2*	24	12	18	10	16	26
A26	1*	10	6	12	4	12	2
A27	78	96	98	100	98	100	98
A28	8	40	44	32	32	44	30
A29	60	96	90	92	94	94	82
A30	34	56	50	36	41	32	26
A31	16	20	28	18	16	22	20
A32	18	54	42	56	48	46	48
A33	14	40	32	60	52	38	38
A34	4	36	34	44	26	22	24
A35	14	48	36	24	44	20	12
A36	46	72	76	68	66	62	62
A37	0*	1*	0*	1*	0*	0*	1*
A38	98	100†	100†	100†	100†	98†	98
A39	1*	4	2	4	2	2	2
A40	0*	4	2	2	2	2	0
B1	92	100	100	100	100	100	100
B2	100	100	100	100	100	100	100
B3	100	100	100	100	100	100	100
B4	100	100	100	100	100	100	100
B5	100	100	100	100	100	100	100
B6	100	100	100	100	100	100	100
B7	100	100	100	100	100	100	100

No	Default DASH	0.27; 73;56	0.27; 73;61	0.27; 73;51	0.27; 60; 63	0.25; 35; 86	0.25; 46; 62
B8	100	100	100	100	100	100	100
B9	100	100	100	100	100	100	100
B10	100	100	100	100	100	100	100
B11	100	100	100	100	100	100	100
B12	96	100	78	100	78	96	100
B13	100	100	100	100	100	100	100
B14	100	100	100	100	100	100	100
B15	66	98	96	86	86	88	78
B16	100	100	100	100	100	100	100
B17	100	100	100	100	100	100	100
B18	70	100	98	98	96	94	92
B19	100	100	100	100	92	100	100
B20	100	100	100	100	100	100	100
B21	44	60	74	84	56	52	56
B22	100	100	100	100	100	100	100
B23	98	100	100	100	100	100	100
B24	96	98	98	96	98	100	98
B25	100	100	100	100	100	100	100
B26	84	98	98	100	96	90	90
B27	44	78	82	66	74	78	70
B28	100	100	100	100	100	100	100
B29	92	100	100	100	100	100	100
B30	64	98	96	92	100	88	90
B31	58	50	74	66	66	54	54
B32	100	100	100	100	100	100	100
B33	100	100	100	100	100	100	100
B34	50	100	100	98	96	98	96
B35	14	48	54	44	40	30	48
B36	4	12	6	12	4	6	6
B37	12	30	38	22	16	26	20
B38	36	76	60	56	66	70	60
B39	4	14	24	22	12	14	14
B40	8	26	18	26	30	18	16
B41	98	100	100	98	100	100	98
B42	20	44	56	42	46	34	34
B43	12	32	28	16	22	22	32
B44	8	48	52	26	36	32	26
B45	14	54	40	52	68	48	56
B46	4	70	60	50	60	58	50
B47	14	54	64	58	64	62	62
B48	4*	12	8	14	4	4	8
B49	0*	1*	0*	0*	0*	0*	0*
B50	0*	0*	0*	1*	2	1*	0*
B51	1*	6	6	20	10	14	10
B52	0*	18	18	10	8	12	16
B53	2	22	38	44	46	36	26
B54	0*	1*	14	2	2	8	2
B55	4	90	64	96	96	96	90
B56	0*	0*	0*	0*	0*	0*	0*
B57	0*	0*	0*	0*	0*	0*	0*
B58	78	100	96	100	100	100	96
B59	0*	0*	0*	0*	0*	0*	0*
B60	0*	1*	0*	1	1*	6	1*
B61	0*	0*	0*	0*	0*	0*	0*

Table 5.2 A summary of the SRs achieved with the six aggressive SA parameter configurations, against the FDS, based on the 500 SA runs. The SR given in brackets are based on the 100 SA runs as given in Table 5.2. IF FDS = improvement factor calculated against the FDS; IF DoF≥14= improvement factor based on compounds with DoF≥14. The IFs have been calculated based on the results given in both Table 5.1 and 5.2. Note the cases which were not solved with the default DASH parameters (*i.e.* 0% SR), but have been solved with one or more of the aggressive parameter settings; their best improvement factor is technically infinite, but here has been capped at 100; when SR achieved with different number of SA moves were compared, the IF was calculated based on the number of SA steps required to reach a solution..

No	Default DASH	0.27; 73;56	0.27; 73;61	0.27; 73;51	0.27; 60; 63	0.25; 35; 86	0.25; 46; 62
A37	0	(1)	1.6	(1)	0.8	0.4	(1)
A40	0.2	(4)	(2)	(2)	(2)	(2)	0.2
B49	0	(1)	0.6	0	0.4	0.4	0.2
B50	0.2	0.4	0.6	(1)	(2)	(1)	0.2
B52	9.4	(18)	(18)	(10)	(8)	(12)	(16)
B54	2	(1)	(14)	(2)	(2)	(8)	(2)
B56	0	0	0	0	0	0	0
B57	0	0.4	0	0.2	0.4	0	0
B59	0	0.2	0.4	0.2	0.2	0.6	0
B60	0.4	(1)	3.4	(1)	(1)	(6)	(1)
B61	0	0	0	0	0	0	0
IF FDS	n/a	7.44	6.13	6.34	7.26	6.87	4.22
IF DoF≥14	n/a	13.67	11.05	11.47	13.32	12.54	7.30

5.4 Discussion

The various criteria used in assessing DASH performance discussed below in order to establish the best performing SA parameters configuration and its improvement factor.

5.4.1 Success rate improvements

All six configurations showed an excellent overall SR improvement, demonstrated by their improvement factors. It is important to note than when considering the FDS, structures which had a SR of 100% under baseline conditions were also included in improvement factor calculations, even though no improvement was possible, thus reducing the overall average value. When only compounds with DoF > 14 are considered, the average improvement factor nearly doubles for all configurations. These latter values are considered to be more relevant measure of the improvement gain, especially given the aims of the thesis. It should, of course, be remembered that this average can be considered to be somewhat skewed by the inclusion of the (capped) factor of 100 improvement assigned to those structures that solved under aggressive settings whilst failing to solve under defaults.

An alternative way of looking at the results is to consider the average SR (rather than improvement factor) of each subset of the FDS, calculated for all six configurations (Table 5.3). Presenting the improvements in this way avoids the issue of calculating a SR

'improvement' when the baseline SR is zero. That said, it lessens the scientific significance of a transition from zero SR (*i.e*. no solution) to a non-zero SR (*i.e*. a solved crystal structure). It is also notable that the improvements observed over the training and test sets are very similar, providing reassurance that the performance of the configurations is not biased toward instances used during the optimisation *i.e*. training set.

Table 5.3 The average SR of the structural complexity groups. The values given have been based on the SRs achieved with up to the 100 SA runs. Please note that the average SRs were calculated by omitting all entries of 100% success. † The standard deviations have been omitted as were found to be irrelevant (as discusses in Chapter 4).

Configuration	Sum of SRs	Median	SR Average† FDS (%)	SR Average † Training Set (%)	SR Average† Test Set (%)
Default	5589	58	32.67	33.26	32.28
0.27; 73;56	6924	96	51.82	53.04	50.98
0.27; 73;61	6936	96	52.00	53.5	50.95
0.27; 73;51	6831	98	50.46	51.18	49.95
0.27; 60; 63	6730	86	50.44	48.54	51.78
0.25; 35; 86	6677	88	48.19	47.93	48.38
0.25; 46; 62	6602	90	48.56	50.25	47.38

Of particular interest are all compounds which, during the baseline calculations, required 500 SA runs in order to solve i.e. A40, B50, B52, B54 and B60. All of these returned a solution within the first 100 SA runs with the aggressive configurations, a remarkable improvement in performance. Furthermore, four of the six previously unsolved compounds (A37, B49, B57 and B59) now returned a solution, with only B56 and B61 remaining unsolved.

A close inspection of the best aggressive DASH solution for B56 (profile χ^2 value of 126 vs. Pawley $\chi^2 = 43$) shows that it is actually a potential solution. The crystal structure overlay (performed with Mercury) confirms that the correct positions of all of the fragments have been found (Figure 5.1); the orientation of the carvedilol molecule is also accurate, and its adopted conformation is close to the reference one, apart from the flipped Bz-O-Me end group. For the purpose of this work however, the solution was considered not to meet the strict solution criteria and was classified as a failure. Nevertheless, the best solution is undoubtedly a significant improvement on the best baseline solution (profile $\chi^2 = 210$) which did not display any chemical sense (Figure 5.2).

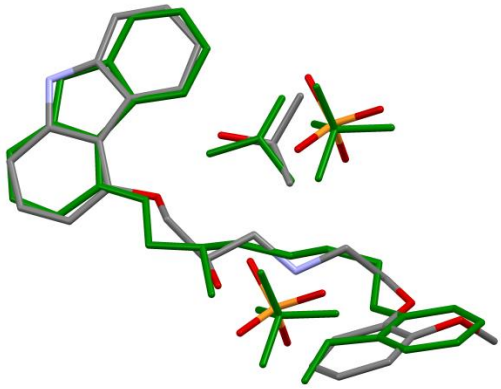


Figure 5.1 Overlay of the best aggressive DASH solution and the reference crystal structure (green)

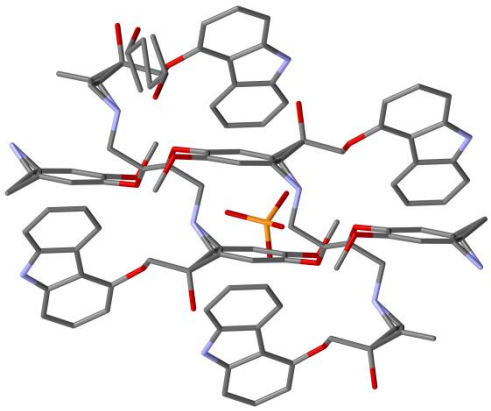


Figure 5.2 The best baseline DASH solution of B56.

The crystal structure of B61 still remains unsolved. Its chances of successful crystal structure solution are further discussed in Section 5.4.4.

Considering the results of Table 5.3, two configurations 0.27/73/56 and 0.27/73/61 were selected as best performing. Their improvements, however, were very similar and did not allow one to be chosen as best performing. As such, the number of highest SRs achieved oved the FDS of the two configurations were taken into account (highlighted in Table 5.1 and Table 5.2; summarised in Table 5.4) and the 0.27/73/56 SA parameter set was selected for future DASH usage.

Table 5.4 A summary of the performance improvements of the 0.27; 73; 56 and 0.27; 73; 61 configuration. 1st, 2nd and 3rd denote the rank of SR improvement, whist the 'Total' gives the value of improvements observed (the sum of top 3 ranked).

0.27; 73; 56					0.27; 73; 61				
1st	2nd	3rd	Total	Reduced	1st	2nd	3rd	Total	Reduced
27	16	12	55	2	27	15	4	46	2

The number of instances where a small reduction in the SR was observed with the aggressive parameters was considered insignificant (*e.g.* A19 default SR = 12, aggressive SR = 10; B31 default SR = 58, aggressive SR = 50).

5.4.2 Gains in calculation times

In an industrial context, the time taken to achieve a crystal structure solution is important; a quick solution will allow further developments that are based on the 3D crystal structure to take place. Thus far, the increase in SR has been used as the primary measure of the improved DASH performance, but here we consider the reduction of overall calculation times and the way in which this is achieved.

1) **Total number of SA runs required**. The increased SR achieved with the aggressive parameters inevitably implies that fewer SA runs are needed in order to return sufficient successful solutions to conclude that the crystal structure has been solved. This reduces the required CPU time for crystal structure determination. Obvious examples include A25, A26, A39, B48 and B51, for which 100 SA runs were needed with the default DASH settings, and only 50 SA runs with the aggressive configurations. Even with compounds solvable within the initial 50 SA runs on defaults, improvements are still seen with the aggressive parameters (e.g. B44-B47, Table 5.5).

   Table 5.5 Default and aggressive SRs achieved for compounds B44-B47.

No	Default DASH	0.27; 73;56	0.27; 73;61
B44	8	48	52
B45	14	54	40
B46	4	70	60
B47	14	54	64

   Taking B44 as an example, the number of runs needed to guarantee a likely structure solution is reduced by a factor of five.

2) **The number of SA steps**. A number of complex crystal structures, such as the aforementioned A40, B52, B54, and B60, have seen a significant reduction in the number of steps required to achieve a solution. Taking ornidazole (A40) as an example, a solution on defaults was only achieved when 5×10^7 steps were employed. In contrast, with the aggressive parameters, 1×10^7 SA steps proved sufficient. The combination of this reduction

and the decreased number of SA runs cuts the solution time from 29 days to only 14 hours of CPU time.

3) **The number of steps required to reach a pre-defined χ^2 value (the χ^2 multiplier)**. Thus far, all DASH calculations have been carried out with the χ^2 multiplier set to 1, ensuring that all of the set number of SA moves are always performed. In practice, setting this multiplier to a higher value (e.g. 3-5) can further reduce the calculation times, by engaging a simplex minimisation when $\chi^2_{SA} \leq$ Multiplier $\times \chi^2_{Pawley}$. As employing the aggressive settings means that lower values of χ^2_{SA} are more likely to be achieved earlier (as a direct result of the high cooling rate) then one might expect this to lead to substantial gains in time. This hypothesis was tested for a χ^2 multiplier of 5 using both default and aggressive settings against three test structures.

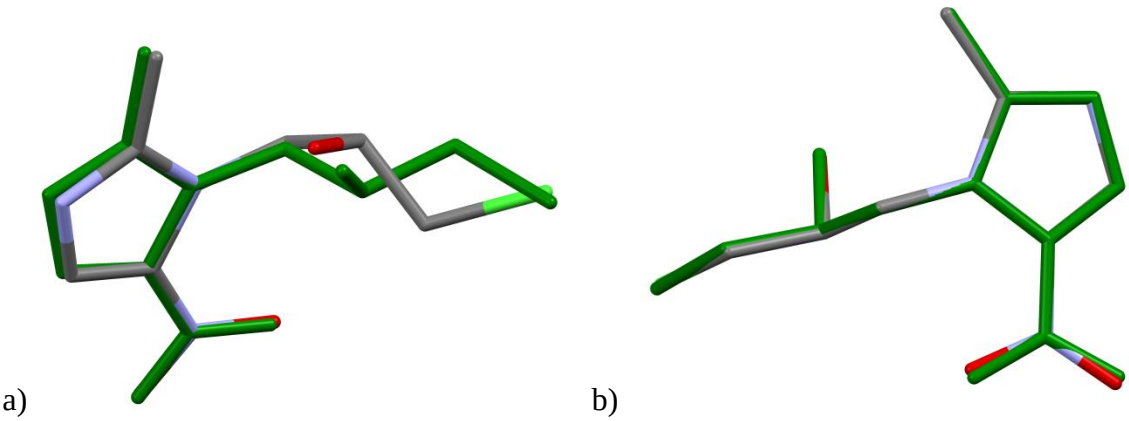
Table 5.6 Calculation times for B45-B47 as a function of the χ^2 multiplier = 5.

	Default SR (%)	Default DASH CPU time	Aggressive SR (%)	Aggressive DASH CPU time
B45	14	4hrs 50 mins	54	3 hrs 24 min
B46	4	4hrs 31 mins	70	3 hrs 18 min
B47	14	2hrs 14 mins	54	1 hr 22 mins

Clearly the improved SRs are maintained and gains in speed are realised.

5.4.3 Crystal structure solution accuracy

The accuracy of the crystal structure solution is another factor of importance. As expected, with the aggressive settings the best DASH solutions retained their good agreement with the reference structures and a small number of structures saw an improvement in the accuracy. In the case of B60, a reduction in profile χ^2 from 5.94 to 5.35 resulted in improvement in the RMSD value from 0.498 Å to 0.225 Å. An even more pronounced improvement was observed for A40 where the reduction in the profile χ^2 from 74 to 30 manifested itself in a significantly better overlay with the reference crystal structure (RMSD reduction from 0.296 Å to 0.075 Å) (Figure 5.3).



a) b)

Figure 5.3 Crystal stucture overlay of the reference A40 crystal ctructure (in green) and: a) the best baseline DASH solution (Default RMSD = 0.296 Å); b) the best aggressive DASH solution (Aggressive RMSD = 0.075 Å). For clarity only one ot the three ornidazole molecules are shown, which is representative of the GoF for all three molecules in the assymetric unit cell. Hydrogen atoms have been omitted.

B56, discussed in Section 5.4.1 can also be considered here, as the accuracy of the best aggressive DASH solution is markedly superior to the best baseline DASH solution, as clearly demonstrated by Figure 5.1 and Figure 5.2.

5.4.4 Statistical Analysis

An ELO analysis was performed for the 0.27/73/56 configuration, following the same approach used in the baseline DASH analysis.

First the ELO calculations were performed against the total DoF, resulting in the correlation given in Equation 5.1 and Figure 5.4 (R^2 factor of 51.7).

$$ELO = ln((SR + 0.5)/(100 - SR + 0.5)) = 7.013 - 0.329 \times DoF_{total} \quad \text{Equation 5.1}$$

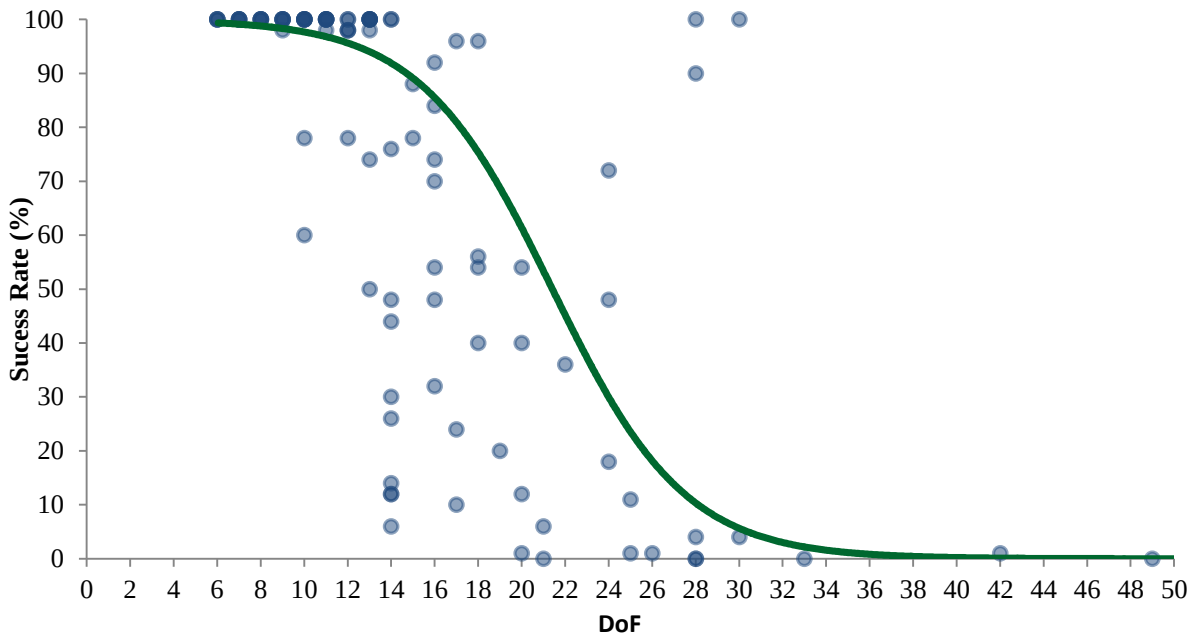


Figure 5.4 The aggressive ELO model based on the total DoF (solid green line). The observed aggressive SRs are shown in blue with increasing opacity with increased number of examples.

The model, just as with the one of the baseline DASH performance, overestimates the actual SRs, but nevertheless describes the general trend in SR reduction. A comparison of the ELO models of the default and aggressive DASH parameters, based on the total DoF, clearly demonstrates the 'shift' in SRs (Figure 5.5). Especially worth nothing, is that based on this model, the SR achieved with the aggressive settings is expected to fall below 1% only after 35 DoF. Of course, this model is only a general description of the SR trend and better estimates were expected with the more flexible ELO model, which take into account the different components of the DoF.

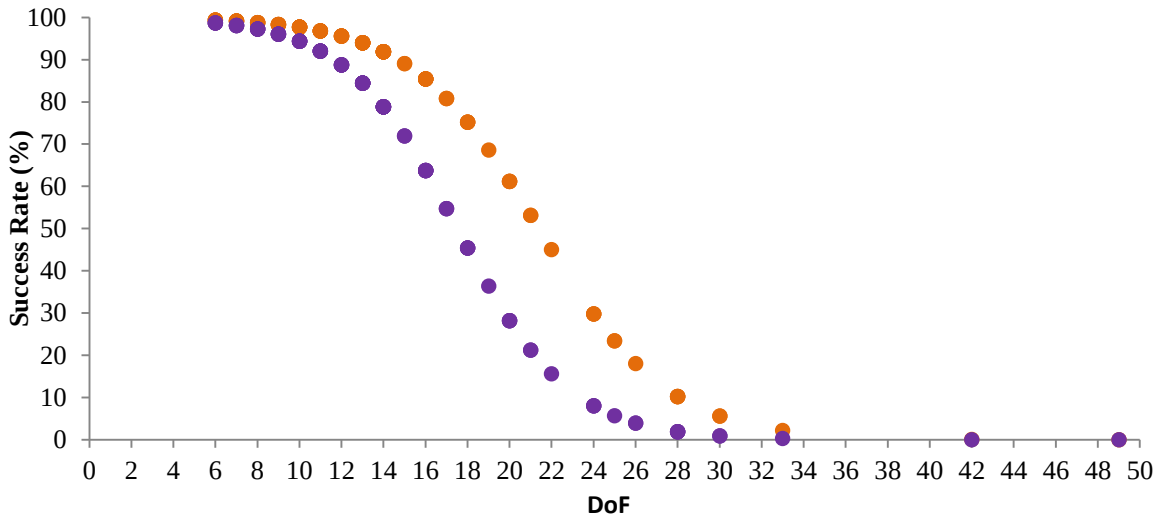


Figure 5.5 Comparison of the default (purple) and aggressive (orange) ELO models based on the total DoF.

The ELO analysis carried out against all three components of the DoF showed, as expected, the orientational DoF to be insignificant in the description of the SR model (Table 5.7) and as such these were excluded in the final ELO calculations.

Table 5.7 Regression analysis of the aggressive ELO *vs.* positional, orientational and torsional DoF.

Term	F-value	p-value
Regression	39.86	0
Positional DoF	12.94	0
Torsional DoF	87.18	0
Orientational DoF	1.01	0.317

Finally, the ELO model of the SR change as a function of the number of positional and torsional DoF is described by Equation 5.2 and its calculated SR pattern is illustrated in Figure 5.6 (R^2 factor of 54.75).

$$ELO = ln((SR + 0.5)/(100 - SR + 0.5)$$
$$= 6.471 - 0.3706 \times DoF_{positional} - 0.4148 \times DoF_{torsional} \quad \text{Equation 5.2}$$

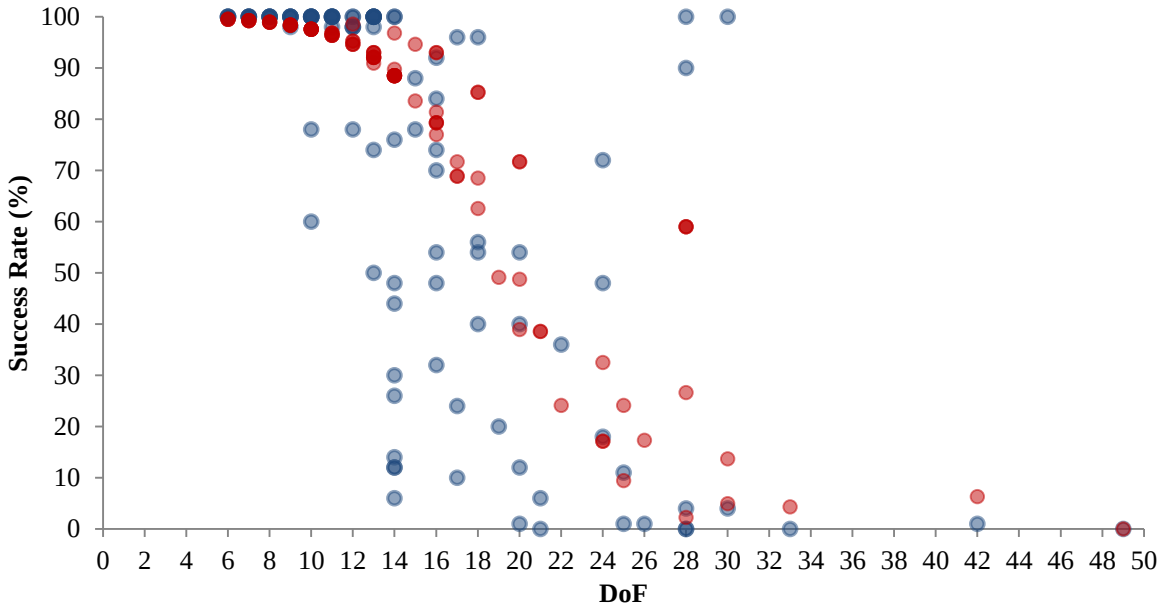


Figure 5.6 The aggressive ELO model based on the positional and torsional DoF (red). The observed SRs are given in blue. The depth of shading of each point is proportional to the number structures representative of a particular DoF.

Again, comparison of the baseline and aggressive DASH parameter models demonstrates the 'shift' in SR performance (Figure 5.7). Of particular interest is the data-point corresponding to 28 DoF (shown in box in Figure 5.7), with a predicted SR of approximately 60%. When all compounds of 28 DoF are considered (Table 5.8), it can be seen that the aggressive ELO model does a better job of accounting for the unexpectedly high SRs observed for A38 and B55. Interestingly, it also reflects the increased difficulty encountered when solving problems that involve larger numbers of torsion angle.

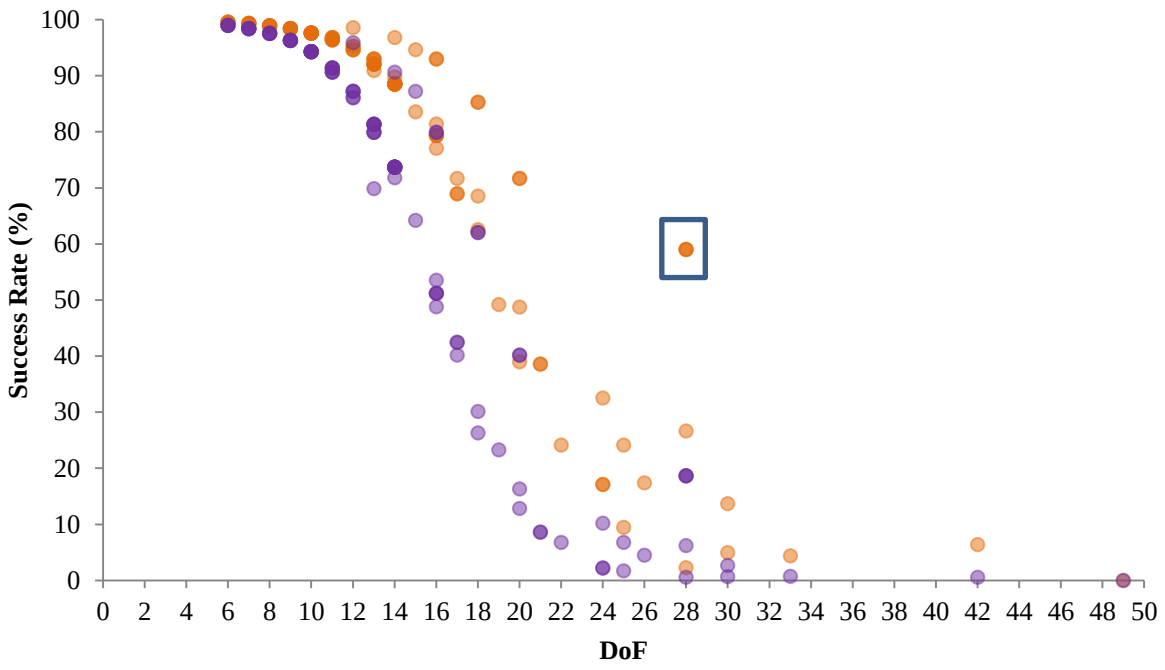


Figure 5.7 Comparison of the default (purple) and aggressive (orange) ELO models based on the positional and torsional DoF. The depth of shading of each point is proportional to the number structures representative of the particular DoF.

Table 5.8 Calculated and experimental SRs of the compounds with 28 DoF. The SRs denoted with * are a result of the 500 SA runs.

Data Set No	DoF $_{total}$	DoF $_{positional}$	DoF $_{torsional}$	$SR_{aggressive\ experimental}$	$SR_{aggressive\ calculated}$
A38	28	12	4	100	59
A39	28	12	4	4	59
B55	28	12	4	90	59
B56	28	9	10	0 *	27
B57	28	3	22	0.4 *	2

The large difference between the predicted and observed SR of A38 and A39 is also notable; it is not obvious why such a remarkable difference in their SR would be observed for two 'identical' problems in terms of their parameterisation. Similar to A29 and A30 (discussed in Chapter 3), both A38 and A39 crystallise in the same space group ($P\bar{1}$), were Pawley-fitted to the comparable resolution (minimum *d*-spacing) and had a similar number of extracted intensities. The only obvious difference is that the data of A38 were collected with synchrotron radiation (λ=0.51561Å), which whilst beneficial (particularly the improvements in angular resolution), is unlikely to be the only factor able to account for such a difference in SR.

Revisiting the crystal structure of B61, which still remained unsolved, the ELO model predicted that the chance of success with the best aggressive configuration are increased by a factor of almost 10, though this still equates to 330 days of CPU time for this particular problem. Whilst this is a significant reduction relative to the predictions for the default SA settings, it was not possible to check this prediction in the time-frame of this work.

5.5 Conclusions

Significant overall improvements in SR were observed with the six selected SA parameters configurations; two sets (0.27/73/56 and 0.27/73/61) were particularly effective and one (0.27/73/56) was selected for future use in DASH, due to its best overall improvement in SR. The solution of four previously intractable crystal structures (A37, B49, B57 and B59) was also achieved with the selected 0.27/73/56 configuration.

The significance of these results lies in the fact that this remarkable improvement in performance has been achieved merely by adjusting the SA control parameter values, with no changes to the underlying SA algorithm. The contribution of irace (López-Ibáñez *et al.*, 2011) in deriving the aggressive SA parameter configurations cannot be underestimated. It is highly unlikely that a set of control parameters which included such a high cooling rate would have been considered for the manual selection. Importantly, the selected aggressive configuration (and in fact any of the six aggressive configurations) can be utilised immediately by manually entering the appropriate parameter values in DASH (in the 'SA options' window). Options of its integration in the core SA code are also currently being considered by the CCDC[14].

5.6 References

López-Ibáñez M, Dubois-Lacoste J, Stützle T and Birattari M (2011) The irace package, Iterated Race for Automatic Algorithm Configuration. Université libre de Bruxelles, IRIDIA, TR/IRIDIA/2011-004, Belgium, http://iridia.ulb.ac.be/irace/, Oct 2015

[14] The CCDC is the current distributor and developer of the DASH software.

6 Exploiting prior conformational knowledge

6.1 Introduction

Global optimisation (GO) based methods of crystal structure determination use a significant amount of prior chemical knowledge explicitly; Well-defined bond lengths and angles are typically held rigid in global optimisation and so the method incorporates prior knowledge implicitly. Furthermore, certain parts of molecules (for example cyclic assemblies) are often treated as rigid. Typically, however, rotatable bonds are allowed to flex, thus conformational space is treated as a continuum rather than a sequence of isolated conformations. However, the conformation of the molecule (as defined by the variable torsion) is rarely utilised[15], regardless of its availability, and as such each of these torsion angles is allowed to vary freely in the range of 0° - 360°.

The use of prior conformational knowledge as constraints during the crystal structure solution process has been previously recognised (as discussed in Chapter 1) and has found particular utility in macromolecular crystallography. For example, a protein molecule from a known crystal structure is often used as a starting 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; Cole *et al.*, 2014; Florence *et al.*, 2005; Middleton *et al.*, 2002). Generally, however, this evidence base is not strong, consisting of isolated examples and lacking quantitative assessment of any gains that are to be achieved. With increasingly more complex crystal structures being of academic and commercial interest, it is therefore timely to re-visit the potential of exploiting conformational knowledge in a more systematic and wide-ranging study, to see if it can build upon the improvements reported in Chapter 5.

Of particular interest is the easily accessible[16] conformational information obtainable from the *ca.* 750,000 crystal structures deposited in the Cambridge Structural Database (CSD). All the tools necessary to search, retrieve and analyse the structures from which relevant molecular geometry information is derived are provided with the Cambridge Structural Database System (CSDS). This information is potentially exploitable by any SDPD software that can incorporate it. In the specific case of DASH, each of the variable torsion angles can

[15] Usually except double and triple bonds. However, in this work all bonds were allowed to rotate freely
[16] In the sense that there is no need to perform additional practical experiments *e.g.* SS-NMR

be described in four ways prior to invoking the SA process:

a) a torsion angle is fully flexible (*i.e*. allowed to vary in the range 0° - 360°)

b) the torsion angle is allowed to vary in the range 0° - 360°, but sampling of the angular space is somehow biased

c) a torsion angle is varied only within a confined, user-specified [17] range

d) an explicit single value for a torsion angle can be entered

If the exact values of all variable torsion angles are known in advance, then the problem of solving the crystal structure is reduced to one of only finding the position and orientation of the molecule in the unit cell, a computationally much simpler problem. The exact conformation of the molecule under study is, however, rarely known in advance and the less restrictive options (b) and (c) are more suitable for introducing prior conformational knowledge. Methods of utilising the CSD-derived information in this way are provided by the Mogul and Mogul distribution bias (MDB) approaches, both of which have already been implemented in DASH and distributed as part of the CSD system package. The differences between these two approaches, and a further novel approach to introducing conformational knowledge, are discussed below.

6.1.1 Mogul and Mogul distribution bias (MDB)

Mogul is a knowledge-based library of molecular geometries derived from the CSD. It derives information on intramolecular geometric patterns from CSD entries and displays the results as histograms with a figure of merit and additional statistics. Taking the C_6-C_5-O_2-C_{20} torsion angle of verapamil HCl as an example, the Mogul distribution (based on *ca*. 11,100 CSD deposited crystal structures) shows that this angle is most likely to adopt a value in the range 0˚ and 20˚ (populated by *ca.* 10,500 of the entries) (Figure 6.1).

[17] Either manually or automatically introduced

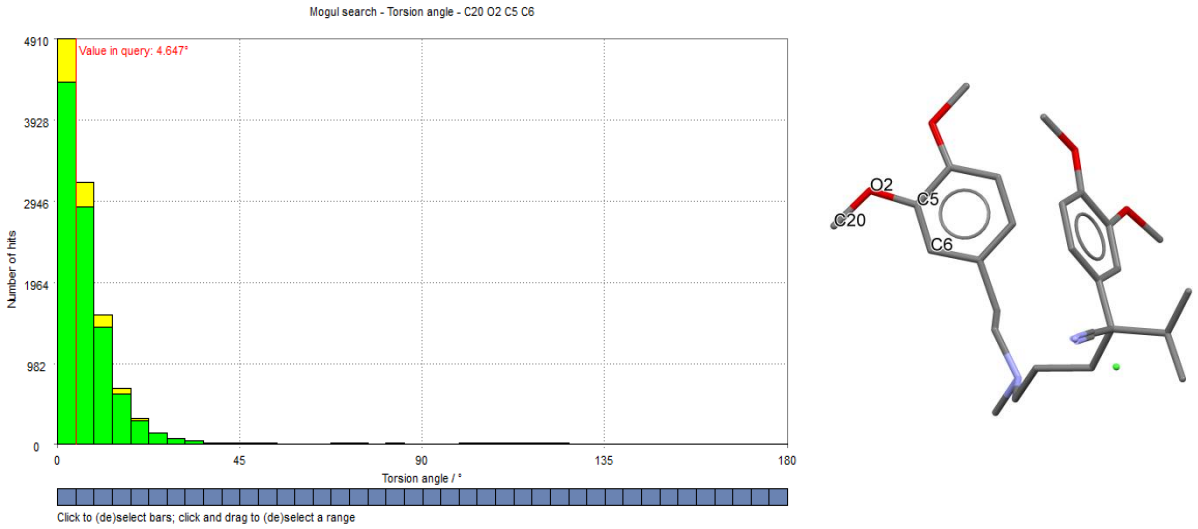


Figure 6.1 The Mogul-derived distribution of the C6-C5-O2-C20 torsion angle based on the CSD entries.

The information contained in the distribution can be utilised in two ways within DASH:

1) **Mogul**. During the SA, the most likely torsion angle ranges from a series of discrete constraints. For example, in the case of the C_6-C_5-O_2-C_{20} torsion angle of verapamil HCl, only values in the ranges of 0˚ to +20˚ and -20˚ to 0˚ are permitted[18] (Figure 6.2). Such a reduction in search space of one torsion angle is not expected to have a notable impact on the overall SR of verapamil HCl, but if similar Mogul-derived restrictions are applied across all of the 14 variable torsions in the molecule, the total search space remaining is closer to that occupied by only 5 freely rotating torsion angles. This approach has been shown to result in a notable improvement of the SR (CCDC; Florence *et al.*, 2005).

[18] Note that for this particular example, the two ranges are adjacent and so effectively form a single range spanning -20° to +20°, restricting the methyl group to be either 20° above or below the plane of the benzene ring

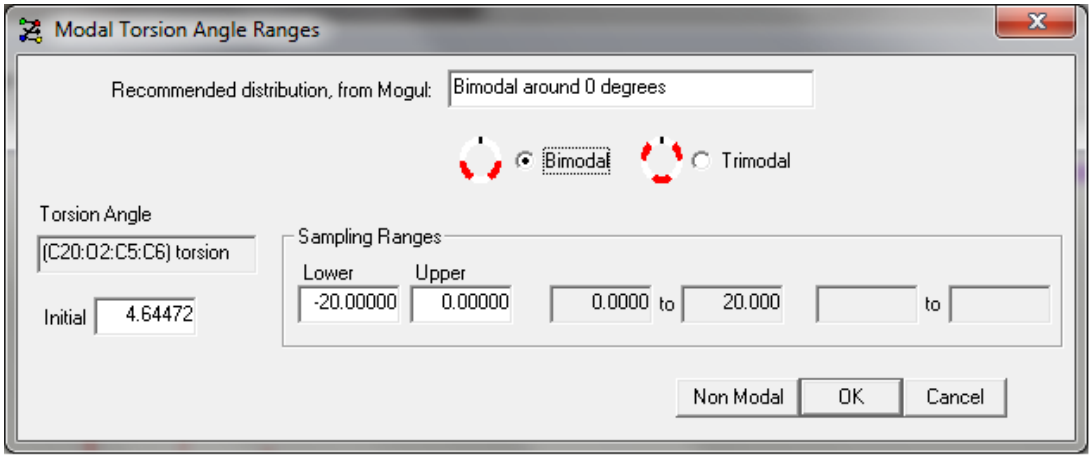


Figure 6.2. The modal ranges of constraint/s as applied by Mogul. Here the user has the option to change the values of the ranges.

2) **MDB**. MDB is a complementary method of exploiting the Mogul-derived conformational information. Here, the Mogul distribution obtained for a torsion angle is used to bias the SA sampling towards torsion angle values that lie within that probability distribution. The feature which clearly distinguishes MDB from the Mogul approach is that here the sampling is more continuous than discrete. During the SA, the full 0° - 360° range is still sampled, but the probability of a value being selected is scaled according to its frequency of observation in the Mogul distribution. Considering again the C_6-C_5-O_2-C_{20} torsion angle, the MDB is applied in the following form:

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 is given first, followed by the instruction to use MDB, the minimum and maximum angular values, the number of bins in the histogram and finally the number of observations in the bin. Therefore during the SA calculation, the probability of sampling a torsion angle value in the first bin is approximately 576 times higher than that of sampling a value in the last bin.

The two approaches therefore introduce the same underlying information in subtly different ways, resulting in the different exploration of χ^2 space during the SA. Importantly, the DASH-Mogul interface is already in place, allowing a rapid and straightforward transfer of the Mogul-derived conformational data into DASH. (Figure 6.3). The use of MDB is fully automated, whilst Mogul permits some user intervention.

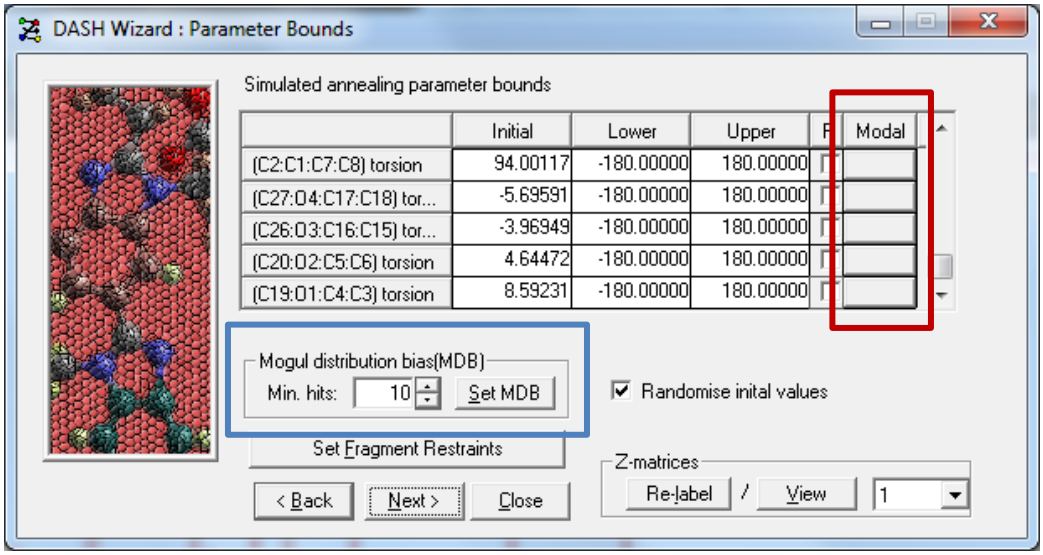


Figure 6.3 The 'parameter bounds' window of DASH. Mogul constraints are applied as modal distributions for each of the torsion angles, with the use of the "Modal" option.

6.1.2 Likely conformers as constrained starting models

The generation of likely 3D conformers plays an important role in the area of drug design (Davies and Richards, 2002; Good *et al.*, 2001; Murray and Cato, 1999; Musafia and Senderowitz, 2010) and as such numerous conformer generation packages have been developed, including VConf (Chang and Gilson, 2003), Confab (O'Boyle *et al.*, 2011), OMEGA (Hawkins and Nicholls, 2012), and Frog (Leite *et al.*, 2007). Each adopts a distinct approach to conformer generation, and many attempt to identify low-energy conformers as those most likely to be observed in nature.

In this work, a CSD-based conformer generator (0.9.3) was employed to output an ensemble of likely conformers based on observed crystal structures, rather than on energy. The conformer generator has been developed at the CCDC and is becoming part of the academic CSD system (November 2015 release). The process starts with a 3D representation of the molecule (Figure 6.4), the bond lengths and angles of which are optionally optimised against likely values taken from Mogul using gradient-based minimisation. The molecule is then split into a number of components connected by rotatable bonds, and a CSD distribution derived for each. Conformers are generated based on these CSD distributions, and the final set of *n* conformers is filtered down using a clustering algorithm which eliminates similar conformations with respect to their distance matrix RMSD. By default, up to 200 conformers are generated, with the option for this value to be defined by the user.

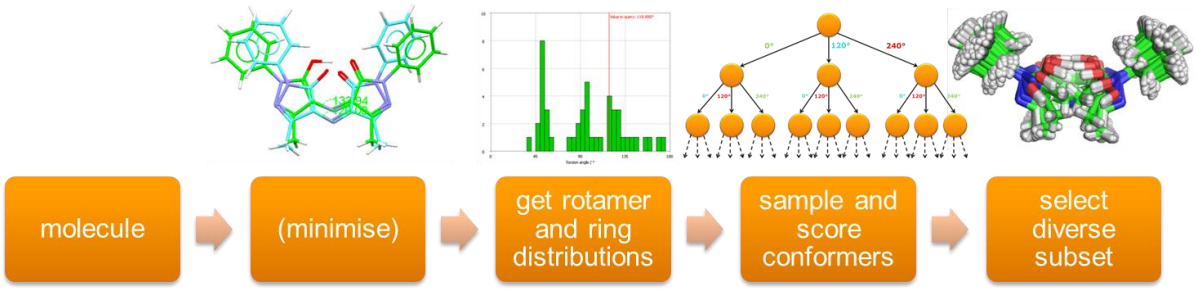


Figure 6.4 Conformer generator (0.9.3) workflow (CCDC, 2015)

Following the generation of the likely conformers, each one is used as a semi-rigid[19] starting model in the SA. A representation of the work-flow created to utilise these conformers is given in Figure 6.5. If one assumes that the conformation of a given conformer is close to that observed in the crystal structure to be solved, a successful solution should be attainable with a short number (relative to the normal fully flexible model) of SA runs each of which performs a fraction of the 10^7 SA moves usually employed. As such, this approach is expected to result in gains in calculation time, rather than increase in the SR, which constituted the primary measure of improvement in previous chapters.

[19] The definition of 'semi-rigid' is that each torsion angle is constrained during the SA to lie ±*x*° from its conformer value, where *x* is a small value *e.g.* 20. This is discussed further in Section 6.1.2.

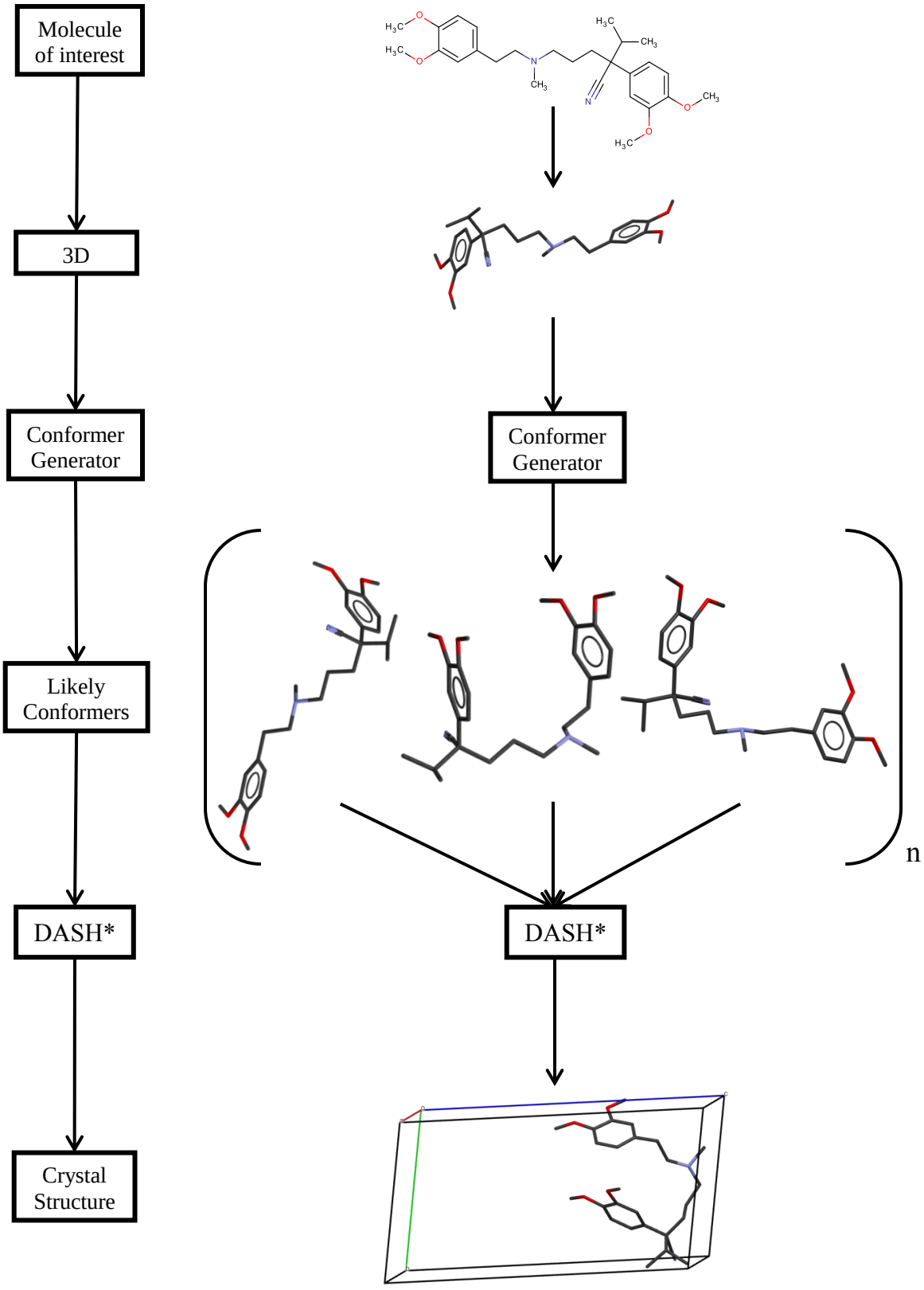


Figure 6.5 The use of likely conformers as constrained starting models to solve crystal structures from powder diffraction data with DASH. (*) denotes that constrains were applied during the SA.

6.2 Experimental

6.2.1 Mogul and MDB

Compounds with a default SR lower than 60% were selected for the evaluation of Mogul and MDB. Following the methodology used with the baseline and aggressive DASH calculations, 50 and 100 SA runs of 1×10^7 moves were initially performed, followed by 500 SA runs of 5×10^7 moves with compounds for which no successful solution was observed in the initial 100 SA runs. The χ^2 multiplier was set to one, and the values of the random seeds were selected as described for the baseline DASH runs (Chapter 3). The option to change any torsion ranges derived by Mogul was not used.

6.2.2 Likely conformers as constrained starting models

Ten compounds were selected for the evaluation of this approach: 4 representatives of crystal structures which display very low SRs with 50 and 100 SA runs (A25, A34, B53 and B54); and the 6 compounds for which a solution was not reached with the default SA parameters (Table 6.1).

200 conformers were generated for each of the 10 compounds, with the use of the CSD Conformer Generator 0.9.3, using its default settings. The input ".mol2" files required were generated from the reference crystal structures with the use of Mogul.

The execution of the DASH runs, of each conformer ensemble, was facilitated by a command-line driven python script – FDASH (Spillman, 2014). FDASH starts by dividing the ensemble of 200 into individual conformers, which are converted into z-matrices with the use of the *makezmatrix.exe* program supplied as part of DASH. A DASH batch file (dbf) is then created for each of the z-matrices, with a user defined number of SA runs and moves (see Table 6.1 for the number of SA runs/steps per conformer).

Table 6.1 FDASH runs performed with the selected 10 compounds.

No	SA runs per conformer	SA moves
A25	2	5×10^5
	3	3×10^5
	5	5×10^5
A34	2	5×10^5
	3	3×10^5
	5	5×10^5
B53	2	5×10^5
	5	5×10^5
	5	5×10^6
B49	2	5×10^6
B54	2	5×10^6
A37	5	8×10^6
B56	5	8×10^6
B57	5	8×10^6
B59	5	8×10^6
B61	5	8×10^6

All of the above DASH runs were performed with a $\pm20^o$ allowed angular range around each conformer torsion angle value, a χ^2 multiplier of 1 and the usual simplex calculation at the end of each SA run. The 'fine control of torsion angle tolerance' option in FDASH, which allows different angular ranges to be input for individual torsion angles, was not used. Additional fragments of the asymmetric unit cell (*e.g.* counter-ions and solvates) were treated in the same way as in normal DASH runs.

6.3 Results

6.3.1 Mogul and MDB

The results of the 50 and 100 SA runs are given in Table 6.2, whilst Table 6.3 summarises the outcomes of the 500 SA runs. Both tables also include information on the accuracy of the Mogul-derived information. Of the 453 variable torsion angles present in the molecules studied;

(a) 309 were constrained to ranges that spanned the values observed in the reference crystal structures

(b) 49 were constrained to ranges which did not span the values observed in the reference crystal structures.

(c) 95 could not be constrained based on CSD observations (see Section 6.1.2 for further discussion).

Table 6.2 Mogul and MDB results based on the 50 and 100 SA runs using default and aggressive parameters. SRs denoted § are a result of MDB constraints based on 7 CSD entries, rather than the usual 10, due to insufficient CSD entries. † denotes the SR achieved with 100 SA runs. $N_{correct}$= number of torsions where observed torsion angle values are within the constrained range; $N_{incorrect}$= number of torsions where observed torsion angle values are outside the constrained range; N_{no_recom}= number of torsions for which no recommendation was made. N_{filter} = number of torsion angles filtered by Mogul. Note that the average improvement factor does not include the Table 6.3 values. The improvement factor for aggressive versus default settings (13.7) was discussed in Chapter 5.

No	DoF_{Total}	$DoF_{torsion}$	$N_{correct}$	$N_{incorrect}$	N_{no_recom}	N_{filter}	Default SR(%)	$Mogul_{Default}$ SR(%)	$MDB_{Default}$ SR(%)	Aggressive SR(%)	$Mogul_{Aggressive}$ SR(%)	$MDB_{Aggressive}$ SR(%)
A7	10	4	3	0	1	0	48	86	100	78	98	100
A16	13	7	2	2	0	3	42	76	60	74	92	94
A18	14	8	7	0	0	1	4	20	30	6	24	32
A19	14	8	4	0	2	2	14	36	68	12	18	44
A20	15	9	6	0	3	0	34	68	70	88	92	88
A21	15	6	5	1	0	0	56	72	52	78	96	86
A22	16	10	8	2	0	0	28	74	82	74	94	98
A23	16	7	6	0	1	0	54	84	78	92	98	94
A24	16	4	2	1	1	0	50	30	26	84	44	48
A25	17	11	7	1	2	1	2†	4	12	24	4	6
A26	17	11	8	2	0	1	1†	8	4	10	30	2
A28	18	3	2	0	1	0	8	36	28	40	54	48
A30	18	6	4	0	1	1	34	20	18	56	56	20
A31	18	12	6	0	3	3	16	6	6	20	16	12
A32	20	8	8	0	0	0	18	38	48	54	74	70
A33	20	5	3	1	1	0	14	24	44	40	62	84
A34	22	14	12	1	1	0	4	26	18	36	60	28
A35	24	6	0	1	1	4	14	2	14	48	16	70
A36	24	12	5	5	2	0	46	56	82	72	80	98
A37	25	13	7	0	3	3	0	0	1†	1†	0†	0†
A39	28	4	1	3	0	0	1†	2†	0†	4	2†	4
A40	30	12	12	0	0	0	0†	0†	0†	4	0†	0†
B21	10	4	0	0	2	2	44	90	76§	60	96	96§
B27	12	3	3	0	0	0	44	92	88	78	100	100
B31	13	7	3	1	3	0	58	92	92	50	98	96
B34	14	8	6	0	2	0	50	72	94	100	100	100

No	DoF_{Total}	$DoF_{torsion}$	$N_{correct}$	$N_{incorrect}$	N_{no_recom}	N_{filter}	Default SR(%)	$Mogul_{Default}$ SR(%)	$MDB_{Default}$ SR(%)	Aggressive SR(%)	$Mogul_{Aggressive}$ SR(%)	$MDB_{Aggressive}$ SR(%)
B35	14	8	8	0	0	0	14	100	92	48	100	100
B36	14	8	7	0	0	1	4	30	44	12	84	44
B37	14	8	7	0	0	1	12	100	96	30	100	100
B38	14	8	7	0	0	1	36	100	98	76	100	100
B39	14	8	7	0	1	0	4	94	68	14	98	96
B40	14	8	7	0	1	0	8	100	90	26	100	96
B42	14	5	1	2	2	0	20	2	6	44	42	6
B43	16	7	4	2	1	0	12	1†	22	32	92	86
B44	16	7	1	0	2	4	8	46	32	48	56	66
B45	16	4	3	1	0	0	14	4	8	54	56	46
B46	16	4	3	1	0	0	4	12	16	70	80	60
B47	18	9	5	1	3	0	14	42	46	54	94	98
B48	20	8	5	3	0	0	4†	2	4	12	4	10
B49	20	14	11	1	2	0	0†	1	2	1†	1†	14
B50	21	6	3	1	1	1	0†	4	4	0†	10	8
B51	21	6	3	1	1	1	1†	62	18	6	86	78
B52	24	6	4	0	1	1	0	28	14	18	68	56
B53	25	13	7	2	0	4	2	0†	8	22	0†	46
B54	26	14	7	1	0	6	0	1	4	1†	4	8
B55	28	4	0	4	0	0	4	6	4	90	78	90
B56	28	10	2	3	0	5	0†	0†	0†	0†	0†	0†
B57	28	22	19	2	1	0	0	0†	0†	0†	0†	1†
B59	33	15	15	0	0	0	0	0†	0†	0†	0†	0†
B60	42	6	6	0	0	0	0	0†	0†	1†	0†	0†
B61	49	43	37	3	3	0	0	0†	0†	0†	0†	0†
Average improvement factor								8.4	9.3		6.6	6.2

Table 6.3 Mogul and MDB results based on the 500 SA runs using default and aggressive parameters. SRs in brackets are a result of up to 100 SA runs as given in Table 6.2

No	DoF_{Total}	$DoF_{torsion}$	$N_{correct}$	$N_{incorrect}$	N_{no_recom}	N_{filter}	Default SR(%)	$Mogul_{Default}$ SR(%)	$MDB_{Default}$ SR(%)	Aggressive SR(%)	$Mogul_{Aggressive}$ SR(%)	$MDB_{Aggressive}$ SR(%)
A37	25	13	7	0	3	3	0	0	NR	NR	0	0.4
A39	28	4	1	3	0	0	(1)	(2)	1.6	(4)	(2)	(4)
A40	30	12	12	0	0	0	0.2	0	0	(4)	0.2	0.4
B56	28	10	2	3	0	5	0	0	0	0	0.8	0.4
B57	28	22	19	2	1	0	0	0	0	0.4	0	(1)
B59	33	15	15	0	0	0	0	0	0.4	0.4	1.2	2
B60	42	6	6	0	0	0	0.4	0.6	0.8	(1)	4.6	5
B61	49	43	37	3	3	0	0	0	0	0	0	0

6.3.2 Likely conformers

The results from the FDASH runs are summarised in Tables 6.4 and 6.5.

Table 6.4 Results of the conformer ensembles evaluation for compounds requiring up to 100 SA runs (with the use of default SA parameters)

No	SA runs per conformer	SA moves	SA Steps ratio	$SR_{default}$ (%)	N solutions (default)	$SR_{aggressive}$ (%)	N solutions (aggressive)
A25	2	5×10^5	2.5	2†	0	24	0
	3	3×10^5	2.8		0		1
	5	5×10^5	1.0		1		2
A34	2	5×10^5	2.5	4	3	36	2
	3	3×10^5	2.8		1		1
	5	5×10^5	1.0		2		4
B53	2	5×10^5	2.5	2	0	22	0
	5	5×10^5	1.0		0		0
	5	5×10^6	0.1		0		0

Table 6.5 Results of the conformer ensembles evaluation for compounds requiring 500 SA runs (with the use of default SA parameters)

No	SA runs per conformer	SA moves	SA Steps ratio	$SR_{default}$ (%)	N solutions (default)	$SR_{aggressive}$ (%)	N solutions (aggressive)
B49	2	5×10^6	12.5	0	2	1†	2
B54	2	5×10^6	12.5	2	6	1†	25
A37	5	8×10^6	3.1	0	0	1†	0
B56	5	8×10^6	3.1	0	1	0	2
B57	5	8×10^6	3.1	0	0	0.4	0
B59	5	8×10^6	3.1	0	1	0.2	4
B61	5	8×10^6	3.1	0	0	0	0

6.4 Discussion

There are several ways in which conformational information can be utilised in a crystal structure determination and here we distinguish been its use "in-process" and its use "*a priori*". In the *a priori* scenario, a molecule is folded to a particular conformation that remains (largely) unchanged throughout a structure determination run. For the in-process scenario, the conformational information is used actively during the structure determination run, to influence or restrict the conformations adopted.

A variety of ways exists for determining a likely conformation *a priori*. For example, a SS-NMR experiment may return intramolecular distances for use as constraints; the CSD can be mined for an observed conformation in the solid state; computational chemistry programs can return the global energy minimum for a molecule (or more likely, a large number of

structures that lies close to the lowest energy calculated). Alternatively, conformations generated during ("in process") a structure determination run can be continually assessed by a number of criteria such as detection of bad intra/intermolecular contacts, isolated molecule energy value and lattice energy. The use of energy as a discriminator during structure determination is attractive, and is based on the observation that molecules in crystal structures generally occupy low-lying areas of the potential energy landscape. However, its use poses two problems. Firstly, the rapid evaluation of the energy of a molecule necessitates the use of force fields, which may not be sufficiently accurate (or well parameterised) for the problem at hand. Higher level calculations can provide much greater accuracy, but at a too severe computational cost. Secondly, the energetic contribution and the diffraction contribution to the overall cost function *(e.g.* χ^2 + P.E.) need to be balanced, even though they are on completely different scales. Solutions to this problem have been proposed and implemented in an SDPD context (Putz, 2001; Putz *et al.*, 1999) but are not, as yet, widely adopted.

The advantage of using conformational information derived from the CSD is that it is based on structures observed in the solid state and can be cast in the form of probability distributions that can be used to influence the values of parameters during a run. Crucially, it does not require modification of the cost function beyond the currently employed, diffraction-based, χ^2 value. The work carried out in this chapter focusses on two "in process" approaches and one *a priori* approach that use CSD-derived conformational information.

6.4.1 Mogul and MDB – 'in-process' biasing

Considerable improvements in SR were observed when the Mogul and MDB approaches outlined above were employed. These gains were obtained regardless of whether the default DASH SA parameters or the aggressive settings were used. Interestingly, the use of constraints that did not include some of torsion angle values seen in the crystal structure did not necessarily preclude obtaining a solution; the simplex minimisation employed at the end of the SA does not take into account the Mogul constraints. For example, with structure A36, 5 of the 12 torsion constraints did not encompass the reference structure values, yet an increase in the SR with both the default and aggressive setting of DASH (and both Mogul and MDB) was still seen. Unsurprisingly, MDB deals better with such cases than Mogul; a MDB distribution does not explicitly preclude a parameter taking an 'unlikely' value (*c.f.* Mogul, where strict bounds are applied that cannot be circumvented during the SA, only in the simplex).

Due to the numerous factors that influence the packing of a crystal structure, it is inevitable that there will be torsion angle values seen in 'new' crystal structures that are not represented in the CSD. Despite the fact that there are in excess of 750,000 crystal structures in the CSD, some torsion angles are only found in a very small number of structures and so their influence on an MDB distribution, or Mogul-derived ranges, is minimal. The torsion angles noted as 'no recommendation' in Table 6.2 are cases where the torsion angle of interest is either poorly represented in the CSD where the torsion angle distribution is nearly uniform. Such cases (around 10% of the total torsion angles of this work) are treated as fully flexible by DASH. In the case of B57, the complex molecule was reported (Bauer *et al.*, 2001) to have an unexpected conformation as a result of a strong hydrogen bonding network (Figure 6.6). It therefore comes as no surprise that the use of conformational information did not allow DASH to solve the structure on default SA settings. However, a solution was obtained with the aggressive DASH setting, with both MBD and Mogul, although the latter required 500 SA runs.

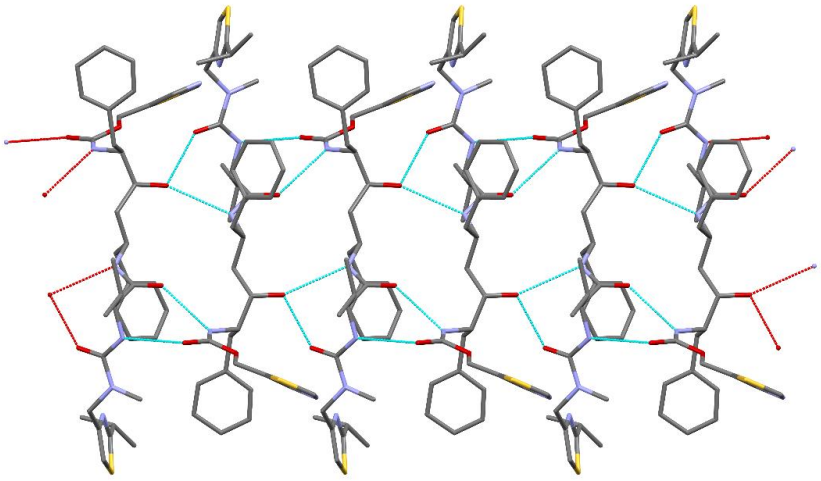


Figure 6.6 Hydrogen bond network of Ritonavir (form II).

Another important factor which must be taken into consideration is the torsion angle definition in the z-matrices. If, for example, the input z-matrix contains a torsion angle that is defined using at least one hydrogen atom, then no distribution is generated[20] and potential Mogul/MDB information is lost. Taking B56 (one of the two remaining unsolved compounds even when the aggressive SA parameters were utilised) as an example, 5 of the 10 torsion angles are described

[20] The angle is 'filtered', on the basis that H-atom positions are often fixed, assumed or otherwise unreliable

in the z-matrix (as automatically generated by *z-matrix.exe*) with the use of a hydrogen atom and as such are ineligible for inclusion the Mogul/MDB distributions. This represents a considerable loss of information, which can be addressed in the future either by the introduction of an option to permit the use filtered results, or by changes to the z-matrix generator, to minimise the use of H-atoms in torsion angle definitions. Despite this loss, two successful solutions were still found with 500 SA runs using MDB.

The reduction in the SR for a small number of additional compounds (*e.g.* B42 and B43 with the default settings and A40, A37 and B60 with the aggressive DASH settings) must be addressed. In the cases of A40 and B60, the DoF are largely positional and orientation (18 out of 30 for A40, 36 out of 42 for B60) and as such they are not heavily influenced by the introduction of the conformational information. Interestingly, even when the correct conformation of A40 is used as input to the SA calculations, DASH fails (on defaults) to solve the structure within 50 SA runs, indicating the extent of the positional/orientational challenge for this particular structure.

Regardless of the above noted limitations, the benefits of employing both Mogul and MDB (with both the default and aggressive DASH settings) are evident in their average improvement factors. Considering that both methods are already implemented in DASH, it is perhaps surprising that these methods are, as yet, not routinely utilised.

Finally, it is worth noting that previously a success rate of 2% was reported A34 using with MDB (Cole *et al.*, 2014) as only 1 of 50 SA runs reached the global minimum (χ^2 = 11). Subsequent analysis has shown that seven other DASH solutions with higher χ^2 values can actually be considered successful, leading to a much higher SR (Figure 6.7).

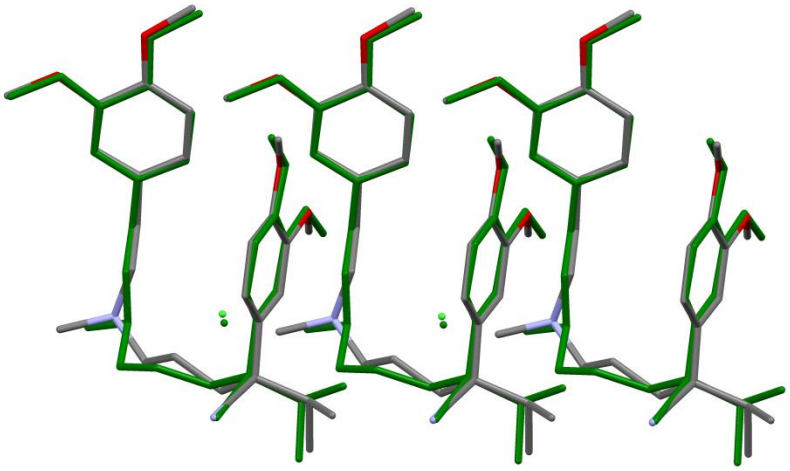


Figure 6.7 An overlay of the A34 reference crystal structure (CURHOM; in green) and a DASH solution of χ^2 of 57. Only 5 of the molecules (including the Cl ions) are presented for simplicity. Hydrogen atoms have been omitted.

6.4.2 Likely conformers as constrained starting models – *'a priori'* approach to introducing conformational information

The results observed with the use of likely conformers as starting models are a direct reflection of the 'accuracy' of the generated conformers. The closer the configuration of a conformer is to that seen in the reference crystal structure, the more likely it is to reach a successful DASH solution with a minimum number of SA runs and moves, and as such yield the largest reductions in CPU time. This suggests that in the ensembles of conformers pertaining to the structures that solved in DASH (A25, A34, B49, B54, B56 and B59), there is at least one conformer which exhibits good agreement with the reference structure. Equally, for those structures that did not solve (A37, B57, B53 and B61) it suggests the ensembles did not contain conformers in sufficiently good agreement with the reference structures. This is confirmed by the findings in Table 6.6 and Figure 6.8.

Table 6.6 Evaluation of the accuracy of the conformer ensembles against the reference crystal structures of the 10 selected compounds. Best RMSD = the RMSD value of the generated conformer which adopts the closes conformation to the reference crystal structure. Conformer rank = the ranking of this 'best' conformer, as output by the conformer generator. † The second best conformer for B56 (4th ranked) has a comparable RMSD of 0.69. (*) The best conformers of A25 and B49 exhibit relatively large variations on all torsions (hence the relatively high RMSD values), but the variations are generally within the $\pm 20^{o}$ torsional flexibility allowed in the FDASH calculations.

Molecules	Solved?	Best RMSD (Å)	Respective Conformer rank	Accuracy
A25	Y	0.95	164	Adequate (*)
A34	Y	0.79	162	Adequate
A37	N	2.03	105	Poor
B49	Y	1.07	126	Adequate (*)
B53	N	1.31	131	Poor
B54	Y	0.52	21	Adequate
B56	Y	0.65	99†	Very Good
B57	N	2.10	192	Poor
B59	Y	0.13	7	Adequate
B61	N	6.24	4	Poor

It is important to note that the RMSD is not the most appropriate metric for this approach; a metric based on comparison of torsion angle values would be more suitable, but was precluded for time reasons. Nevertheless, it gives some sense of the fact that structures where the conformation is close to that observed are more likely to lead to successful solutions. Additionally, it should be noted that in many cases multiple conformers led to successful solutions. For example, B54 returned 25 successful solutions derived from 15 of the 200 starting models.

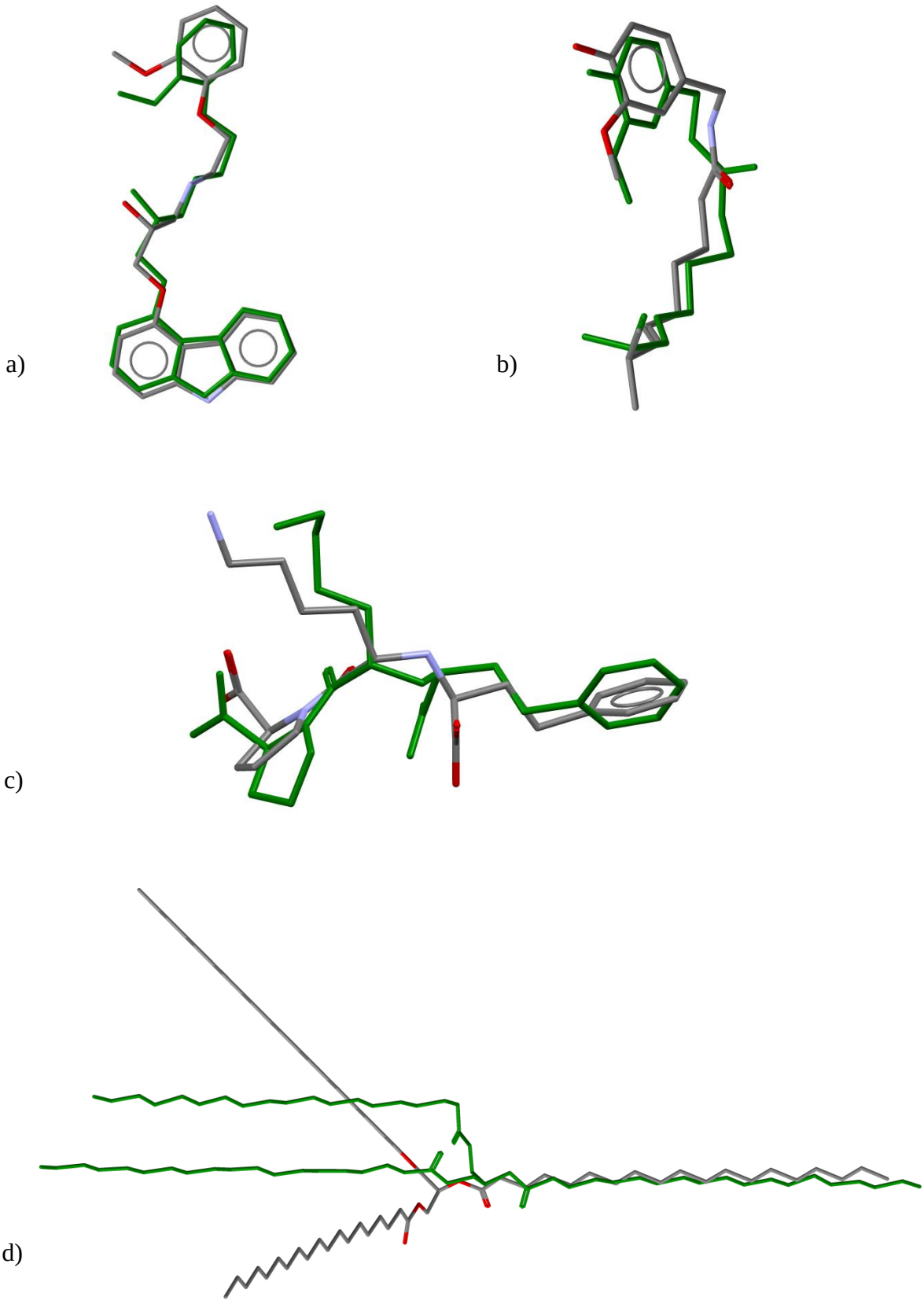


Figure 6.8 Overlays of the reference crystal structures (in green) and the best conformer of: a) B56; b) A25; c) B53 and d) B61 (as in Table 6.6).

All compounds with an RMSD around 1 Å or less result in a successful solution with both default and aggressive SA settings. For compounds with an RMSD >> 1 Å, the structures were intractable given the torsional flexibility and number of SA runs/moves allowed. Although the values of both of these factors could be increased in attempt to achieve successful solutions, the resulting increase in calculation time would defeat the purpose of this approach.

In the case of B56, two of the generated conformers are close to the reference conformation. This is reflected by the fact that the lowest χ^2 value of 85 improves upon previous values obtained from *all* the DASH runs on B56 in this work. Additionally the conformations of the three C_{19} chains of B61 are correct in the best ranked conformer (*i.e*. conformer 1), but due to the poor values generated for the three 'basal' torsions that control their orientations, the overall RMSD is very poor. The same three incorrect values for these vital torsions are seen in all of the 200 conformers of B61 (and in fact with MDB/Mogul); as a result this compound remains unsolved when using prior information.

Beside the accuracy of the input conformation, two additional factors have a significant impact on the chances of success of the FDASH runs:

1) The torsional angle flexibility allowed during the SA.
2) The total number of SA steps performed with for each conformer (given by the number of SA runs × the number of SA steps)

In order to gain an insight into what would be appropriate values for both of the above settings, a number of DASH runs were performed for each on the 10 compounds. First, the reference crystal structure conformation was used as a rigid body starting model (RB; *i.e*. all torsion angles were fixed at their reference values during the SA) for 50 SA runs, each of 10^7 moves, to represent a 'best case' scenario. Secondly, the same model was allowed torsional flexibility of ±10, ±20, and ±30 degrees around each of the torsion angles. For all of these runs, the facility within DASH that allows one to output the χ^2 as a function of number of SA moves was enabled. The resulting "χ^2 vs. moves" plots (see Appendix B) were visually compared to their fully flexible (FF) counterparts in order to establish a torsional flexibility value that would allow sufficient flexibility for a reasonably accurate starting conformer to improve, whilst approximating the performance of the RB calculations (Table 6.7).

Table 6.7 Results of the preliminary DASH runs. All of the presented SRs are a result of 50 SA runs, each performing 1×10^7 SA moves. DEF = DASH runs with the default SA parameter values, AGG = DASH runs with the aggressive SA parameter values. FF = fully flexible; RB = rigid body

	SR (%)									
	RB		±10˚		±20˚		±30˚		FF	
No	DEF	AGG	DEF	AGG	DEF	AGG	DEF	AGG	DEF	AGG
A25	100	100	96	100	96	100	90	98	2	24
A34	100	100	100	100	100	100	100	100	4	36
A37	98	100	80	98	76	100	66	82	0	0
B49	100	100	100	100	100	100	100	100	0	0
B53	100	100	100	100	100	100	100	100	2	22
B54	24	40	28	28	22	44	0	34	0	0
B56	2	14	6	14	0	4	0	2	0	0
B57	100	100	100	100	100	100	100	100	0	0
B59	26	32	6	32	12	28	0	4	0	0
B61	4	12	4	20	8	4	1	0	0	0

In all cases, the use of ±10˚ led to results practically identical with those obtained with the rigid body. As expected, with more flexibility, the SR was reduced and the number of moves required to achieve a solution increased. Nevertheless, even with ±30˚ (*i.e.* a total space of 60˚ around each of the flexible torsions) the results proved advantageous relative to the fully flexible DASH calculations. Based on these results, a value of ±20˚ was chosen as having the best balance between flexibility and performance for subsequent use. However, one cannot ignore the decrease in SR observed with complex/intractable examples, even when a RB starting model is used, indicating the need for a larger number of SA runs per conformer at this level of complexity. For example, B56 solved only once in 50 default SA runs when the RB model was used, which goes some way to explaining why the compound was intractable when no prior conformational information was employed.

In order to establish the 'optimal' number of SA steps (given by the number of SA runs × the number of SA steps) to be allowed when assessing conformers, the plots in Appendix B were examined to establish the lowest number of SA moves required to reach a solution. Unsurprisingly, the results demonstrated that this number of steps varies depending on the complexity of the problem. As such there was no *one* optimal value of SA steps which could be used for all of the 10 compounds during the evaluation of this approach. In the case of A34, the value of 12.5×10^5 moves was arrived at taking into consideration the fact that a speed gain of at least two (over the default performance of DASH) is required.

An important assumption in the use of the conformer approach is that for crystal structures with $Z' > 1$, (such as B54 and B59) the same starting conformer is used for each independent

molecule. This proved to be a valid approach for B54 and B59, and based on the observation that independent molecules in the asymmetric unit generally adopt quite similar conformations (Cruz-Cabeza and Bernstein, 2014; Weng *et al.*, 2008), it should be of general applicability.

Whilst only a small sample (10) of structures was used in the conformer work, the fact that these structures are at the current limits of SDPD means that the six out of ten successes achieved is a good indicator of the potential of this approach.

6.4.3 Statistical analysis

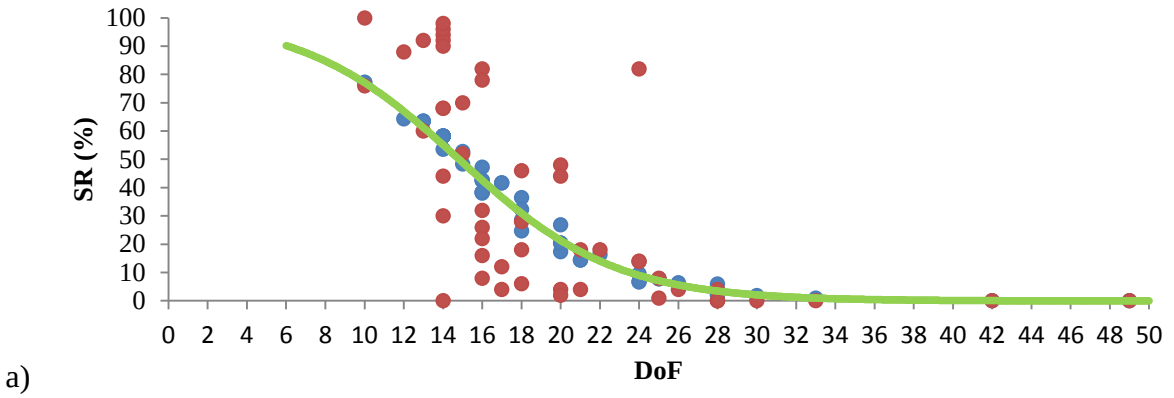
Following the procedure established with the baseline and aggressive DASH calculations, ELO analysis of the results from the 4 combinations of Mogul and MDB used (*i.e.* with the default and aggressive parameters) was performed (Tables 6.8 and 6.9; Figure 6.9). One instantly recognisable difference from the results obtained with previous ELO analysis is the higher significance of orientational DoF in describing the data (all *p* values are much lower than the value of 0.197 returned by the ELO model of default DASH against the FDS). This may be attributable to an increased representation of larger numbers of orientational DoF in the 51 molecule subset and / or the influence of conformational bias reducing the impact of torsional DoF.

Table 6.8 Regression analysis of the default Mogul and MDB ELO models *vs.* positional, orientational and torsional DoF.

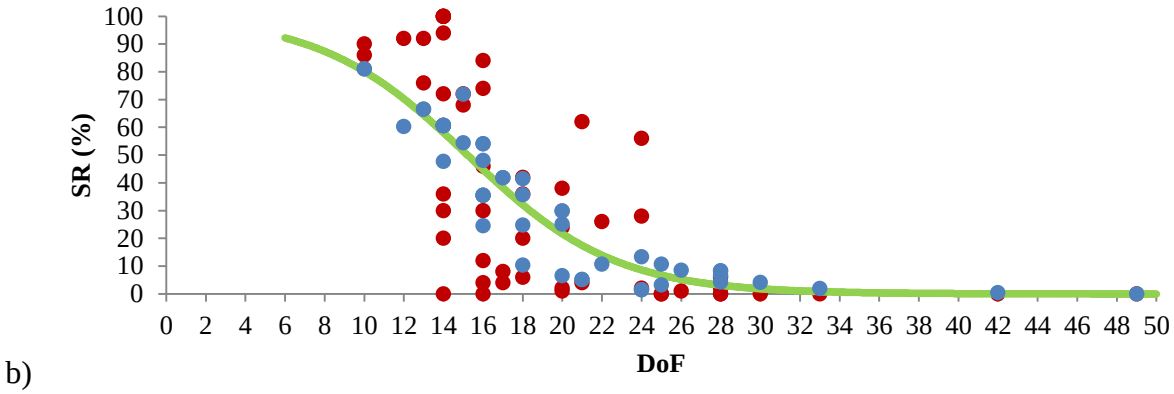
Method	$Mogul_{Default}$		$MBD_{Default}$	
Term	F-value	p-value	F-value	p-value
Regression	11.91	0.000	15.07	0.000
Positional DoF	9.00	0.004	9.94	0.003
Torsional DoF	17.82	0.000	20.43	0.000
Orientational DoF	3.54	0.066	6.21	0.016

Table 6.9 Regression analysis of the aggressive Mogul and MDB ELO models *vs.* positional, orientational and torsional DoF.

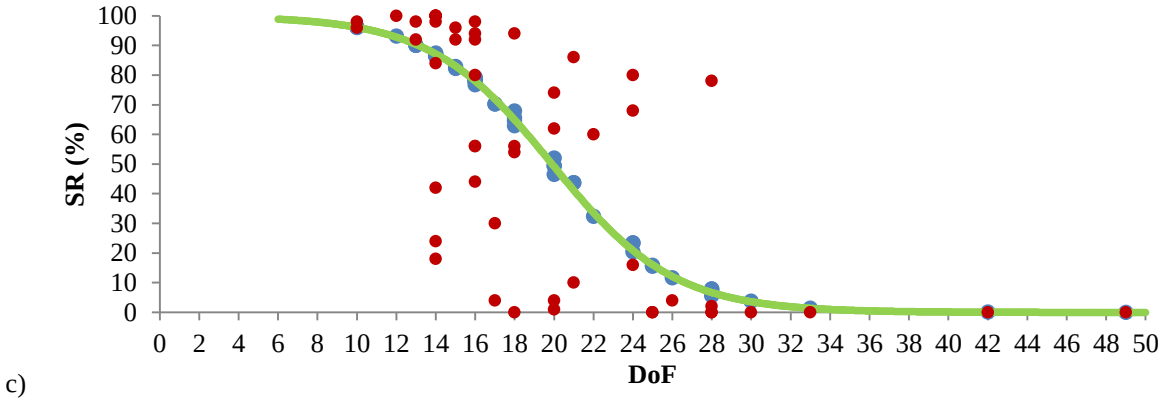
Method	$Mogul_{aggressive}$		$MBD_{aggressive}$	
Term	F-value	p-value	F-value	p-value
Regression	15.87	0.000	12.27	0.000
Positional DoF	7.64	0.008	3.44	0.070
Torsional DoF	29.94	0.000	21.91	0.000
Orientational DoF	5.20	0.027	6.70	0.013



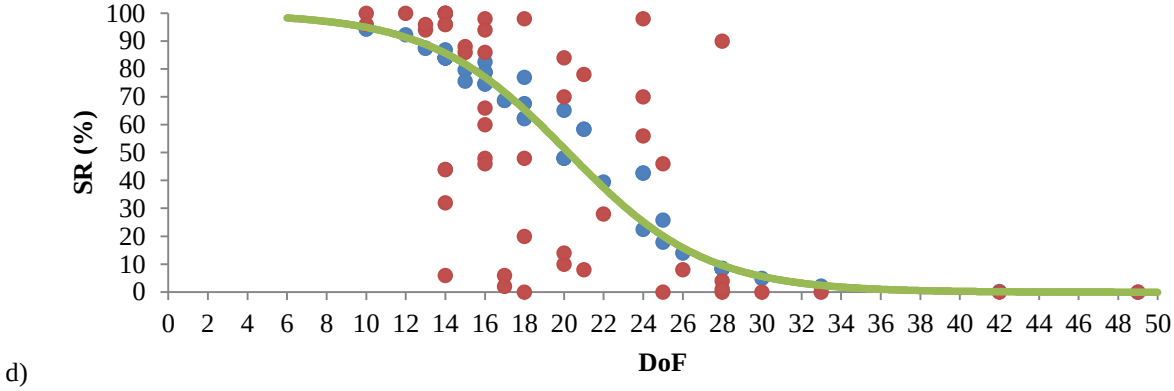
a)



b)



c)



d)

Figure 6.9 ELO models based on the results form a) $Mogul_{Default}$; b) $MDB_{Default}$; c) $Mogul_{Aggressive}$; and d) $MDB_{Aggressive}$. In all figures the observed SR are given in red, the calculated SRs based on the total DoF are shown in green and the calculated SRs based on the appropriate components of the DoF are shown in blue

As expected, the aggressive settings continue to significantly outperform the DASH defaults and the performance of Mogul and MDB is nearly identical (Figure 6.10)

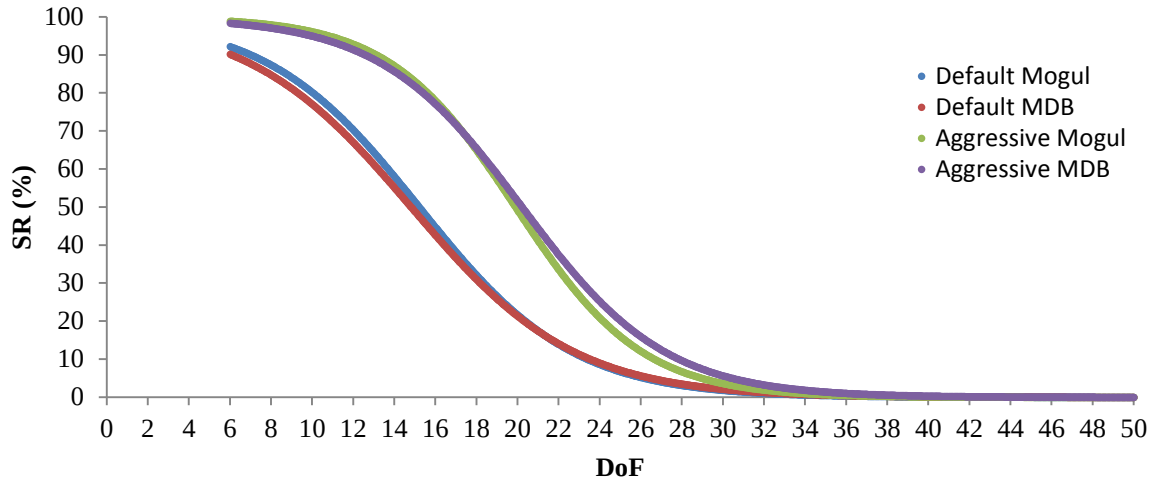


Figure 6.10 An overlay of the four Mogul/MDB ELO models based on the results from the 51 tested compounds

Direct comparison of the 4 ELO models in Figure 6.9 with their "no conformational information" counterparts is inappropriate due to the different number of data points involved. In order to facilitate a direct comparison an assumption was made that the compounds not tested with Mogul/MBD would *at least* retain their SR values when Mogul/MBD are applied. The comparison plot below (Figure 6.11) is based on this assumption. It shows clearly the (almost equal) benefit of inclusion of Mogul or MDB information in the structure determination process, whether DASH is running on defaults or with the aggressive settings. Interestingly, it also shows that use of the aggressive settings alone still outperforms DASH on defaults when used with Mogul or MDB input.

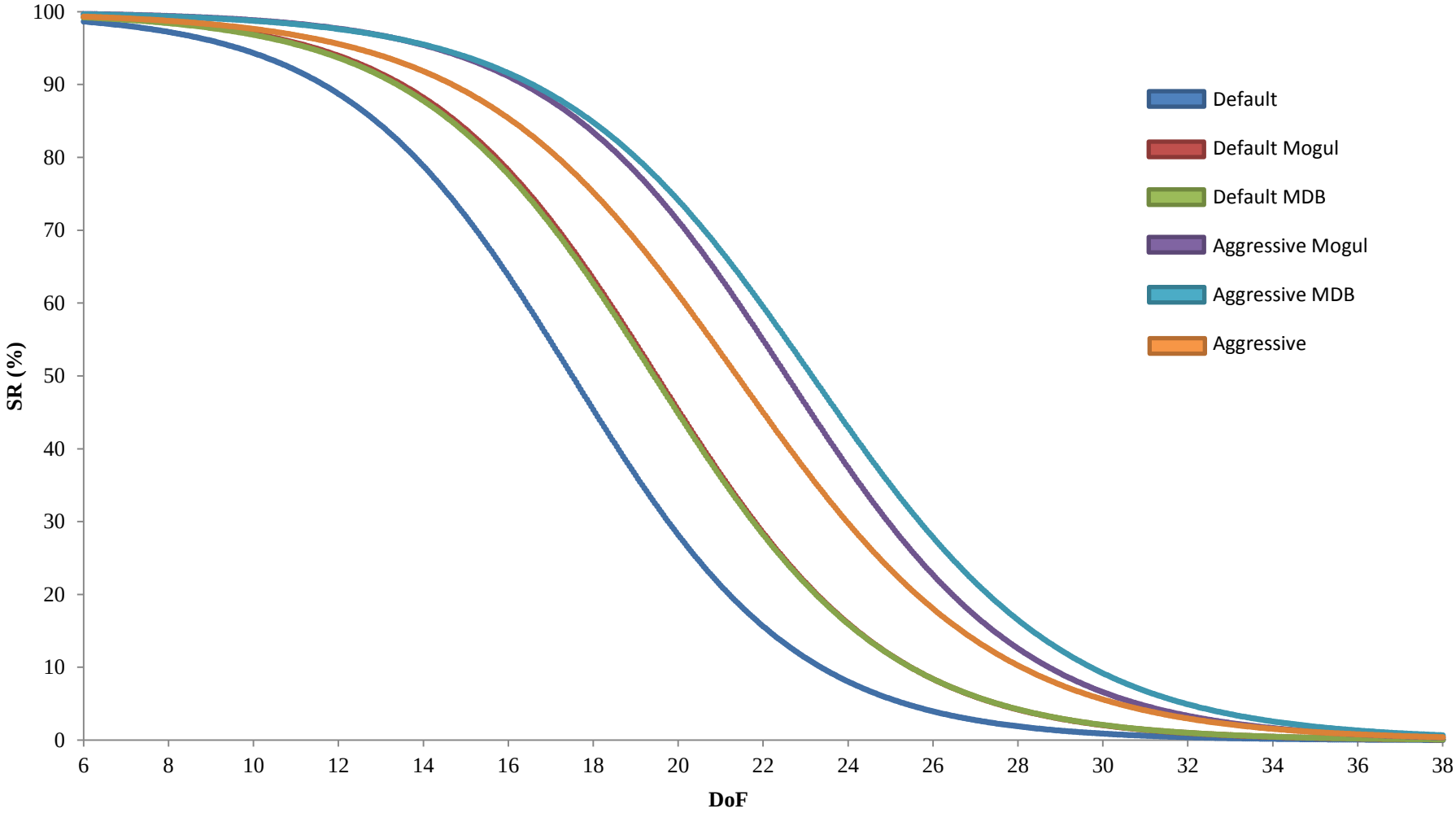


Figure 6.11An overlay of all ELO models based on the total DoF (calculated against the results of the FDS). Please note that whilst the ELO models which take into account the different components of the DoF are considered a better description of the results, their overlay is difficult to interpret and as such the models based on the total DoF are shown. An expanded 6-38 DoF-scale has been used to facilitate comparison of the distributions.

Whilst the use ELO models has its limitations (as noted in previous chapters), it nevertheless presents a more statistically sound basis for comparison of the performance of the six distinct case scenarios (*i.e.* the combinations of default/ aggressive and Mogul/MDB) than the various averages reported in Tables 4.16 and 5.3.

6.5 Conclusions

The work described in this chapter presents a comprehensive study of the effects which including prior conformational knowledge has on SDPD. In particular, the results from MDB and Mogul provide strong evidence that prior conformational information derived from the CSD should be routinely employed when faced with complex problems where the conventional DASH approach has not succeeded. One may actually argue more strongly that the use of conformational information should be the default in DASH; the necessary tools are already in place and the fully automated MDB option is particularly attractive.

There remains, however, work to be done. In particular, the z-matrix definition issue needs to be addressed; the impact of chirality in describing the model and the use rotomer information can also be evaluated. Additionally, it is far from clear that the current SA implementation is the best one in which to integrate prior information. It is entirely possible that CSD-derived conformational information may be harnessed more effectively by other GO-based methods, such as other simulated annealing implementations, or genetic algorithms.

In the case of the generated ensembles of conformers, there was no obvious relationship between success and the rank of the conformer that achieved that success. If a link were to be established, this would allow fewer conformers to be run, improving the performance gains still further.

6.6 References

Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W and Morris J (2001) Ritonavir: An extraordinary example of conformational polymorphism. *Pharm. Res.* **18**:859-866

CCDC Solving the powder pattern of verapamil hydrochloride. Cambridge Crystallographic Data Centre, http://www.ccdc.cam.ac.uk/Lists/ResourceFileList/Verapamil_DASH.pdf, Oct 2015

CCDC (2015) Conformer generator (0.9.3) user manual.

Chang C-E and Gilson MK (2003) Tork: Conformational analysis method for molecules and complexes. *J. Comput. Chem.* **24**:1987-1998

Cole JC, Kabova EA and Shankland K (2014) Utilizing organic and organometallic structural data in powder diffraction. *Powder Diffr.* **29**:S19-S30

Cruz-Cabeza AJ and Bernstein J (2014) Conformational Polymorphism. *Chem. Rev.* **114**:2170-2191

Davies EK and Richards WG (2002) The potential of internet computing for drug discovery. *Drug Discovery Today* **7**:S99-S103

DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, Das D, Vorobiev SM, Iwai H, Pokkuluri PR and Baker D (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473**:540-543

Florence AJ, Shankland N, Shankland K, David WIF, Pidcock E, Xu XL, Johnston A, Kennedy AR, Cox PJ, Evans JSO, Steele G, Cosgrove SD and Frampton CS (2005) Solving molecular crystal structures from laboratory X-ray powder diffraction data with DASH: the state of the art and challenges. *J. Appl. Cryst.* **38**:249-259

Good AC, Mason JS, Green DVS and Leach AR (2001) Pharmacophore-based approaches to combinatorial library design.Combinatorial library design and evaluation: Principles, software tools, and applications in drug discovery

Hawkins PCD and Nicholls A (2012) Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *Journal of Chemical Information and Modeling* **52**:2919-2936

Leite TB, Gomes D, Miteva MA, Chomilier J, Villoutreix BO and Tuffery P (2007) Frog: a FRee Online druG 3D conformation generator. *Nucleic Acids Res.* **35**:W568-W572

Middleton DA, Peng X, Saunders D, Shankland K, David WIF and Markvardsen AJ (2002) Conformational analysis by solid-state NMR and its application to restrained structure determination from powder diffraction data. *Chem. Comm.*:1976-1977

Murray CM and Cato SJ (1999) Design of libraries to explore receptor sites. *J. Chem. Inf. Comput. Sci.* **39**:46-50

Musafia B and Senderowitz H (2010) Biasing conformational ensembles towards bioactive-like conformers for ligand-based drug design. *Expert Opinion on Drug Discovery* **5**:943-959

O'Boyle NM, Vandermeersch T, Flynn CJ, Maguire AR and Hutchison GR (2011) Confab - Systematic generation of diverse low-energy conformers. *Journal of Cheminformatics* **3**:8

Putz H (2001) Structure solution from powder diffraction data by a combined global optimization of several cost functions: Problems and perspectives, in *European Powder Diffraction EPDIC 7* (R. D and E.J. M eds) pp 53-58, Materials Science Forum, Barcelona, Spain

Putz H, Schon JC and Jansen M (1999) Combined method for ab initio structure solution from powder diffraction data. *J. Appl. Cryst.* **32**:864-870

Scapin G (2013) Molecular replacement then and now. *Acta Cryst. Sect. D* **69**:2266-2275

Spillman MJ (2014) Enhancing and accelerating computational methods of crystal structure determination from powder diffraction data, Phd Thesis, University of Reading, Reading

Weng ZF, Motherwell WDS, Allen FH and Cole JM (2008) Conformational variability of molecules in different crystal environments: a database study. *Acta Cryst. Sect. B* **64**:348-362

7 General conclusions

This work has focussed on improving the performance of a global optimisation approach to solving molecular organic crystal structures from X-ray powder diffraction data. An initial survey (Shankland *et al.*, 2013) showed that SDPD was very effective in terms of the structural complexity of problems which could be successfully solved in this way. However, the general trends noted do not capture the difficulties associated with solving structures at the higher end of the complexity scale. In general, as structural complexity (in terms of degrees of freedom to be optimised) increases, so does the number of global optimisation runs needed to deliver a reasonable chance of locating the global minimum. This, combined with the increasing number of function evaluations required for each run, means that the time taken to (possibly) obtain a solution can become a significant deterrent to even embarking on an attempt.

The work described in this thesis has made significant inroads into the above problem, with conclusions validated by testing on more than 100 molecular crystal structures. By tuning the parameters of DASH's simulated annealing algorithm, an approximately ten-fold increase in the success rate (defined as the number of times the global minimum, or points very close to it, is located, relative to the number of runs performed) was obtained. At the upper end of the complexity scale, this enabled solutions to be achieved for previously refractory structures. Exploiting CSD-derived conformational knowledge as part of the SDPD process also resulted in improvements in success rate and was also particularly effective at the upper end of the complexity scale. Importantly, these two distinct approaches are complementary and their combination yields still further gains. There is sufficient evidence to suggest that the adoption of the so-called 'aggressive parameter settings' (CR=0.27, N_1=73, N_2=56) should be immediately adopted by all DASH users and become the default setting in the next release of DASH. Furthermore, the work has shown that the existing DASH implementation for leveraging CSD-derived conformational information is both effective and largely automatic, and that with some changes to the z-matrix definition routine, it is also a strong candidate for adoption as a default setting in DASH. The work performed using conformers as input is extremely promising, given that it was applied to some of the most complex problems, but needs to be investigated on many more compounds before strong recommendations can be made about its general applicability.

Computing power has proven to be a vital component in this work, not only enabling irace to formulate its parameter recommendations, but also enabling adequate number of DASH runs[21] to be performed on a sufficient number crystal structures in order to draw valid conclusions about performance improvements. Regardless of its limitations (as discussed in Chapters 3 and 5), the ELO analysis captured the general improvement trends and allowed ranking of the various approaches based on their performance; MDB+aggressive was identified as best performing, followed by Mogul+aggressive. The use of the aggressive parameters alone outperformed both MDB+default and Mogul+default.

It remains difficult to simply look at a 2D chemical sketch of a structure and conclude from that alone whether its 3D crystal structure can be solved by SDPD - there are too many other factors to consider. Nevertheless, looking at the structures dealt with during this work, one can feel now more confident in not 'ruling out' complex structures at that stage. A better understanding of the distribution of stationary points on the χ^2 agreement hypersurface as a function of structural complexity is needed before one can explain why some structures with a certain number of degrees of freedom present significantly greater challenges than others with the same number of degrees of freedom.

When presented with the task of solving an unknown crystal structure using DASH, the following recommendations can be made:

1. Regardless of apparent or probable structural complexity, the best possible PXRD data should be collected. In the majority of cases, laboratory-based PXRD will be sufficient but the variable count time scheme should always be employed.

2. The best possible 3D input model should be constructed. If a model can be obtained from an existing good-quality crystal structure (e.g. a polymorph of the compound of interest) it should be used. Mogul geometry checks should always be applied prior to use.

3. Time spent obtaining the best possible Pawley fit to data is time well spent. The Pawley χ^2 value is the benchmark against which the SA χ^2 is compared in determining how close to the global minimum the current best structure is. As such, it should be well determined.

[21] The MDASH utility, which enables multiple copies of DASH to be executed simultaneously on multi-core CPUs, was extensively used in this work. The use of MDASH is strongly recommended on any CPU with multiple physical cores.

4. In the structure solution stage, the optimised values of the simulated annealing (CR=0.27, N_1=73, N_2=56) should be assigned. Depending on the complexity of crystal structure under study the following recommendations can be made:

a) For compounds with DoF <14, 50 SA runs each performing 5×10^6 SA moves are sufficient to ensure a high level of certainty that the crystal structure will be determined.

b) Compounds with $14 \leq$ DoF ≤ 20) require 50 SA runs, each consisting of 1×10^7 SA moves to ensure a high level of certainty that the crystal structure will be determined. The use of Mogul or MBD is likely to be beneficial, but is not considered to be strictly necessary with the suggested number of SA runs/steps.

c) If the structure has $21 \leq$ DoF ≤ 27, 50 to 100 SA runs, each consisting of 5×10^7 SA moves are sufficient to ensure a high level of certainty that the crystal structure will be determined, if used in combination with MDB/Mogul constraints.

d) For crystal structures with DoF greater than 27, the recommended start point is 500 SA runs of 5×10^7 SA moves in combination with MDB constraints. The use of likely conformers as constrained starting models is also suitable for this level of complexity. However, due to the currently incomplete evaluation of the method, it is recommended only if the aforementioned 500 SA runs have not reached a successful crystal structure solution. A $\pm20^o$ torsional flexibility is considered appropriate for conformers generated from variety of sources (*i.e* it is not only applicable to CSD-generated conformer ensembles). Furthermore, 5 SA runs per conformer, each of 8×10^6 SA moves should be sufficient to reach a solution, if a conformer closely resembling the adopted crystal structure conformation is present in the ensemble. Alternatively, 10 preliminary SA runs of 5×10^7 moves can be performed in order to evaluate the change of χ^2 observed as a function of the number of SA moves. The point at which no further changes in the χ^2 value are seen can be used as a guide to the number of SA moves to be employed with this approach of introducing prior conformational knowledge.

7.1 References

Shankland K, Spillman MJ, Kabova EA, Edgeley DS and Shankland N (2013) The principles underlying the use of powder diffraction data in solving pharmaceutical crystal structures. *Acta Cryst. Sect. C* **69**:1251-1259