

Space group selection for crystal structure prediction of solvates†

Aurora J. Cruz Cabeza,^a Elna Pidcock,^{*b} Graeme M. Day,^a W. D. Sam Motherwell^b and William Jones^a

Received 9th February 2007, Accepted 14th March 2007

First published as an Advance Article on the web 23rd March 2007

DOI: 10.1039/b702073b

The most populated space groups for a selection of solvates of chiral and achiral molecules with common solvents are presented to assist crystal structure prediction calculations on these complex systems.

Introduction

Many programs for the computer generation of crystal structures (*e.g.* UPACK,¹ Accelrys Polymorph Predictor)² take advantage of space group symmetry to reduce the number of variables in searches of the complex potential energy surface of crystal structures. The generation of crystal structures is, therefore, carried out independently in each space group. As the total number of space groups is considerable (230), a common approach is to generate structures in only the 10 to 15 most frequently observed space groups.^{3–5} Such an approach assumes that the likelihood of finding a low energy crystal structure in a particular space group is related to the proportion of structures in the Cambridge Structural Database (CSD)⁶ that are observed in that space group. Fortunately, a very substantial proportion of structures (91%)⁷ found in the CSD belong to only 15 of the 230 available space groups.

If the system of interest has one crystallographically independent molecule in the asymmetric unit ($Z'' = 1$),[‡] a complete crystal structure prediction (CSP) study in the common space groups can be completed in as little as a few days for a small molecule on a modest single processor machine.⁸ For $Z'' = 2$, however, calculations within the same number of space groups may take more than two to three months of computational time⁹§ as the searchable space becomes a function of six additional variables (*i.e.* position and orientation of the second molecule in the asymmetric unit).¹⁰ In the case of $Z'' = 3$, complete CSP calculations in 10–15 space groups become substantially more demanding: we are only aware of two studies addressing such complex and computationally expensive systems.^{11,12} Multicomponent crystals (*i.e.* cocrystals, hydrates, solvates or salts) will have a minimum of $Z'' = 2$. As CSP calculations are now used to predict the structures of these multicomponent systems and

their likelihood of formation,⁹ an approach to restrict the computational expense of such studies might involve a further reduction in the number of space groups considered.^{13–15} For that purpose, however, we believe that a more specific space group population analysis—beyond the general statistics for all organic crystal structures in the CSD—would be useful, since differences in populations might be significant between different families of crystal structures. Given the chemical nature of the components of a crystal, space groups should be prioritised based on the most relevant statistics.

Solvates¶ constitute a family of crystal structures which has attracted growing interest in materials science. In earlier statistical studies on solvates, Gorbitz and Hersleth¹⁶ and Nangia and Desiraju¹⁷ reported the occurrences of common solvent molecules in crystal structures deposited in the CSD. Furthermore, in the study by Nangia and Desiraju, solvate occurrences were corrected to reflect the different propensities of the various solvent molecules to be incorporated in the crystal lattices during crystal growth. Whilst these studies were mainly focused on the frequency of solvate occurrence, the aim of our present study is to carefully analyse space group distributions in solvates taking into account molecular chirality.

Firstly, we highlight the importance of molecular chirality in the provision of appropriate space group statistics. Occurrences of solvates are then analysed together with chirality: the distributions of chiral (^CM) and achiral (^AM) main components of the solvate are given for the different solvate families. Finally, space group distributions with consideration of chirality (^CM/^AM) and solvent (S) are presented and analysed and their implications for CSP calculations discussed.

Methodology

Structures of interest were retrieved from vs. 5.26 of the CSD (including Nov 04, and Feb 05 updates) using ConQuest.¹⁸ In order to obtain a dataset of structures compatible with the common atom types parameterised in force fields, only neutral organic crystal structures containing subsets of the atoms H, D, C, N, O, S and halogens were allowed. The restriction to only two different chemical residues in the asymmetric unit (RES = 2), excluded heterosolvates (MS₁S₂) and solvates of cocrystals (M₁M₂S), and resulted in an initial data set of

^aThe Pfizer Institute for Pharmaceutical Materials Science, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK, CB2 1EW

^bThe Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK, CB2 1EZ E-mail: pidcock@ccdc.cam.ac.uk; Fax: +44 1223 336033; Tel: +44 1223 762531

† Electronic supplementary information (ESI) available: Numerical data of the space group statistics. See DOI: 10.1039/b702073b

‡ Z'' is the number of crystallographic non-equivalent molecules in the asymmetric unit whereas Z' is the number of formula units in the asymmetric unit. See ref. 13 for a description of the notation and definitions.

§ Although this is generally reduced by the use of distributed or parallel computing resources.

¶ In this case the term solvate is used for molecular crystals containing an organic solvent.

12 579 crystal structures, including two-component cocrystals, solvates and hydrates. The RES = 2 restriction does not exclude different stoichiometries of binary crystals. Only structures which contained the following solvents were further analysed: H₂O, MeOH, EtOH, AcOH, EtOEt, AcOEt, acetone, DMSO, DMF, MeCN, CH₂Cl₂, CHCl₃, CCl₄, dioxane, THF, benzene, toluene, *para*-xylene and hexane (7712 Refcodes). Hydrate structures, of which there were 3792, were excluded from our general analysis as they constitute a very important proportion of the subset (close to 50%). However, hydrate statistics are reported for comparison. Datasets were then processed by an algorithm written to detect the presence of a chiral centre in organic molecules.¹⁹ A flow chart of the methodology is given in Fig. 1.

Space group distributions with chirality

In Table 1, we show space group occurrences for various sets of crystal structures: all structures in the database (CCDC statistics);⁶ structures with two distinct chemical residues per asymmetric unit (RES = 2); and the further subsets of RES = 2 structures for common solvents with an achiral main component (^AM) and chiral main component (^CM). There are clear differences in space group populations between the four sets of structures, especially when chirality is considered (significance levels are higher than 0.001—detailed statistical analysis is presented in the supplementary information).[†]

For all the structures in the CSD, the five most common space groups are $P2_1/c > P\bar{1} > P2_12_12_1 > C2/c > P2_1$, accounting for ~79% of the structures. Although $P2_1/c$ and $P\bar{1}$ remain the two most popular space groups for the RES = 2

Table 1 Space group distributions (%) of various sets of crystal structures. RES is the number of different chemical entities and N the number of structures. Space groups 1–4 are common centrosymmetric space groups whereas 5–8 are common chiral space groups

	Space groups	All ^a	RES = 2	RES = 2 ^b	
				^A M	^C M
Centrosymmetric	$P2_1/c$ (%)	35.2	24.8	35.6	14.4
	$P\bar{1}$ (%)	22.0	22.9	37.2	14.1
	$C2/c$ (%)	7.9	7.4	8.3	3.4
	$Pbca$ (%)	3.6	1.8	1.9	0.8
Sohnke	$P2_12_12_1$ (%)	8.2	13.0	2.8	26.4
	$P2_1$ (%)	5.6	11.5	2.5	25.1
	$P1$ (%)	1.0	2.2	0.7	5.3
	$C2$ (%)	0.9	2.6	0.3	4.6
	Rest (%)	15.6	13.8	10.7	5.9
	N	363217	12579	2361	1559

^a From ref. 7. ^b This study and for crystal structures containing common organic solvents (excluding hydrates). See text for the list of solvents selected.

subset, $P2_1/c$ was found to be significantly less dominant and important changes are observed in the ordering of the remaining space groups. On the other hand, consideration of chirality makes the space group population differences more striking, as chiral molecules, when crystallised from enantiomerically pure solutions, must crystallise in a Sohnke|| space group. Therefore, for solvates with ^CM, enantiopure structures most frequently crystallise in $P2_12_12_1 > P2_1 > P1 > C2$ (we note that for the first time, structures crystallising in $C2$ contribute noticeably) whereas racemic solvates crystallise in $P2_1/c, P\bar{1}$ and $C2/c$. 81% of the solvates with ^AM crystallise in the three most popular space groups ($P\bar{1}, P2_1/c$ and $C2/c$) compared to 65% in the general statistics.

Chirality and popularity of solvents

The general occurrences (O) of different solvent molecules amongst the total 3920 solvate crystal structures presented here are similar to those reported by Nangia and Desiraju,¹⁷ and Gorbitz and Hersleth¹⁶ (Table 2). Our extra contribution is to examine in detail the space group distributions in these families of solvates (see section below), accounting also for the chirality of the main component.

For S = methanol, ethanol and AcOEt, cocrystallisation with ^CM is more frequent than with ^AM: chiral molecules represent approximately 60% of these subsets. The remaining solvents: acetone, DMSO, DMF, MeCN, CH₂Cl₂, CHCl₃, dioxane, benzene and toluene, are observed more frequently with ^AM. The solvent subsets of DMSO, DMF, dioxane, MeCN and CHCl₃, for example, are composed of approximately 70% achiral molecules. In the majority of the cases for ^CM, crystal structures crystallising in Sohnke space groups are more prevalent than crystal structures crystallising in centrosymmetric space groups. Exceptions are observed for the solvent molecules benzene and dioxane, both of which show a greater proportion of racemic structures than enantiopure structures. Interestingly, of all the solvent

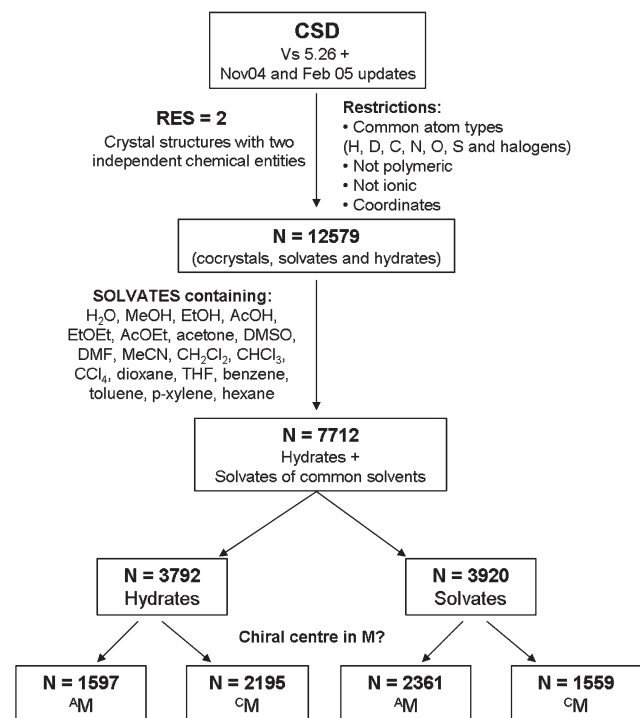


Fig. 1 Flow chart of the methodology followed. N is the number of crystal structures, ^AM and ^CM are main components of the solvate without a chiral centre or with at least one chiral centre, respectively.

|| Sohnke space groups are those without mirror or inversion symmetry.

Table 2 Occurrences (O) and percentages of solvates with and without chiral centres. N is the number of crystal structures and M is the main component of the solvate (^AM for achiral and ^CM for chiral). Only sets with N > 100 structures are shown

S	N	O (%)	M		^C M ^a	
			^A M (%)	^C M (%)	Sohnke (%)	Non Sohnke (%)
Water	3792	—	42.1	57.9	42.6	8.6
Methanol	650	16.4	38.3	61.7	48.5	11.5
Ethanol	257	6.5	42.8	57.2	36.2	15.9
AcOEt	134	3.4	38.1	61.9	46.3	14.9
Acetone	310	7.8	52.9	47.1	28.4	15.8
DMSO	238	6.0	74.4	25.6	16.4	8.4
DMF	185	4.7	79.5	20.5	10.3	9.2
MeCN	231	5.8	69.7	30.3	17.8	11.3
CH ₂ Cl ₂	455	11.5	66.8	33.2	18.2	12.3
CHCl ₃	434	11.0	69.8	30.2	15.2	14.1
Dioxane	131	3.3	80.2	19.8	6.1	13.0
Benzene	354	9.0	63.0	37.0	14.1	18.7
Toluene	127	3.2	75.6	24.4	15.0	7.1

^a ^CM Sohnke is calculated by adding up the percentages of ^CM structures of that particular solvate crystallising in *P*₂*1*₂*1*, *P*₂*1*, *P*₁ and *C*₂. ^CM non-Sohnke is calculated by adding up the percentages of ^CM structures of that particular solvate crystallising in *P*₂*1*/*c*, *P*₁, *C*₂/*c* and *Pbca*.

molecules studied here, only benzene and dioxane possess a centre of symmetry.

While we observed the above differences in solvent popularity for chiral/achiral M, we cannot draw any physical conclusions about the solvents' preferences in solvate formation; any causal interpretation of the observed differences in space group distributions would require more information than is present in the CSD and is therefore out of the scope of the present work.

Space group distributions with chirality and solvent nature

Differences in the distribution of structures over space groups are also important when solvates are divided according to solvent nature. Table 3 compares the distribution of structures over space groups for methanol and DMSO solvates with ^AM and ^CM.

For ^AM, DMSO solvates tend to crystallise noticeably more often in *P*₁ than in *P*₂*1*/*c* (by ~11%). *P*₁ and *P*₂*1*/*c* are also the

Table 3 Distribution of MeOH and DMSO solvates with chiral and achiral M over 8 of the most common space groups. RES is the number of different chemical residues and N the number of structures per subset

Space groups	RES = 2 ^A M		RES = 2 ^C M	
	S = MeOH	S = DMSO	S = MeOH	S = DMSO
<i>P</i> ₂ <i>1</i> / <i>c</i> (%)	39.0	31.6	9.0	15
<i>P</i> ₁ (%)	35.3	42.4	7.0	15
<i>C</i> ₂ / <i>c</i> (%)	5.6	10.7	1.5	3
<i>Pbca</i> (%)	4.4	1.1	1.2	—
<i>P</i> ₂ <i>1</i> ₂ <i>1</i> (%)	1.6	4.0	34.7	26
<i>P</i> ₂ <i>1</i> (%)	3.6	2.3	32.7	23
<i>P</i> ₁ (%)	0.4	—	5.5	3
<i>C</i> ₂ (%)	—	1.1	5.7	12
Rest (%)	10.1	6.8	2.7	3
N	249	177	401	61

most popular space groups for ^AM MeOH solvates. Contributions from *C*₂/*c* are very important in DMSO solvates, but less so for MeOH solvates. In the latter case, the populations of *C*₂/*c*, *Pbca* and *P*₂*1* are very close. For ^CM in enantiopure crystals, *P*₂*1*₂*1* and *P*₂*1* are the most highly populated, while *C*₂ also seems to contribute importantly in the particular case of DMSO solvates. Interestingly, in more than 50% of these DMSO solvates in *C*₂, the DMSO molecule sits on a two-fold axis and shows positional disorder.

We summarise the space group distributions of the most common solvates with ^AM and ^CM in Fig. 2 and 3, respectively. The distributions of the total set of solvates and hydrates are given in the first two columns, followed by the specific subsets. Numerical data is available in the ESI.†

In many of the solvate families where the main component is achiral (^AM), between 70–80% of the crystal structures can be found in the space groups *P*₂*1*/*c* and *P*₁ (Fig. 2). Such cases include solvates of methanol, ethanol, AcOH, AcOEt, DMSO, DMF, MeCN, CHCl₃, dioxane, THF, benzene and toluene. Although in most ^AM solvate subsets *C*₂/*c* is the third most populated space group, variations in its relative importance within the different subsets are evident: while 15% of ^AM acetone solvates crystallise in *C*₂/*c*, only 5.5% do for ^AM ethanol solvates. The fourth most populated space group differs between the families.

^CM solvates in enantiopure form (Fig. 3) are found, with a high probability, in the most common Sohnke space groups *P*₂*1* and *P*₂*1*₂*1*. In the cases of EtOH and AcOEt, approximately 70% of all ^CM structures are found in *P*₂*1* and *P*₂*1*₂*1*. A non-negligible number of solvate structures are also found in other Sohnke space groups such as *P*₁ (toluene [12.9%], acetone [9.6%] and DMF [7.9%]) or *C*₂ (DMSO [11.5%]). On the other hand, for racemic ^CM solvates, the centrosymmetric space groups *P*₂*1*/*c* and *P*₁ are most popular. The overall greatest popularity of *P*₂*1*/*c* and *P*₁ for solvates of dioxane (65.4%) and benzene (42.8%) is due to the greater proportion of racemic structures found (Table 2). Intriguingly, these observations may indicate that the inversion symmetry of dioxane and benzene drives systems to crystallise in space groups with inversion centres. Approximately 50% of benzene and dioxane molecules sit on special positions of the crystal structures. This observation is not only important in terms of space group (those with inversion are preferred) but also of the stoichiometry, as 50% of these solvates are 1 : 1 whereas the remaining 50% are 2 : 1 (M : S). *Pbca*, the sixth most popular space group in the entire CSD, is populated by chiral-molecule containing solvates with a probability of less than 1%.

Implications for crystal structure prediction

From the data presented here, we find that the general space group statistics from the CSD as a whole are not suitable for the particular case of solvates (Table 1): the space group populations in our set of structures with two distinct chemical residues per asymmetric unit (RES = 2) are considerably different from the overall statistics. Furthermore, taking into account (i) the nature of the solvent molecule, S, and (ii) the chirality of the main component, M, may help us make a better

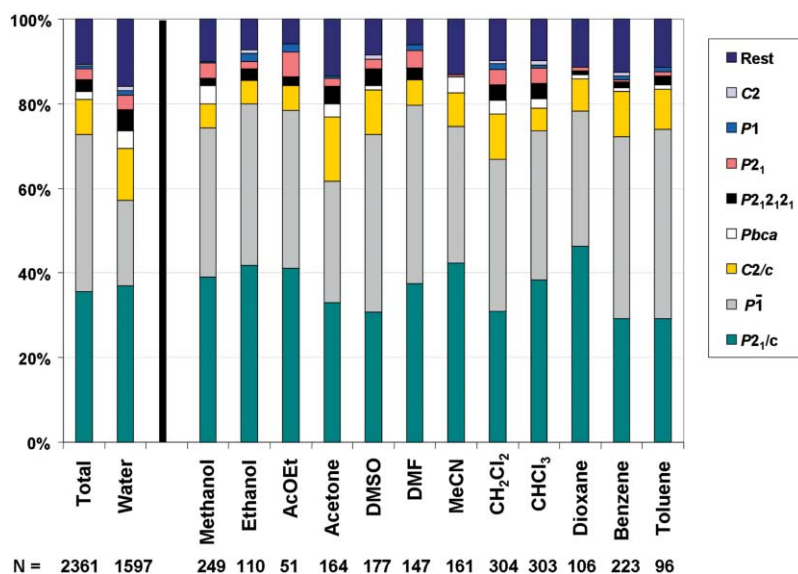


Fig. 2 Summary of the space group frequencies of the total solvates and hydrates (left) and some of the solvate subsets (right) for solvates with $^A M$. N is the number of structures per subset.

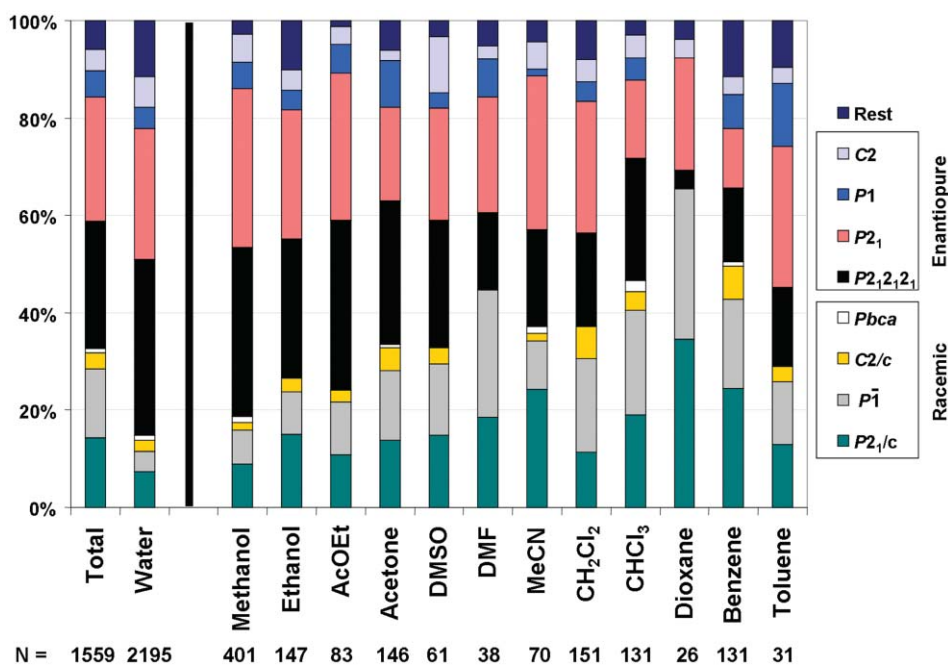


Fig. 3 Summary of the space group frequencies of the total solvates and hydrates (left) and some of the solvate subsets (right) for solvates with $^C M$. N is the number of structures per subset.

judgement of the relative importance of space groups in these crystal systems. In terms of M, three possibilities must be considered: (i) where M is an achiral molecule ($^A M$); (ii) where M is chiral ($^C M$) and one enantiomer is present in the crystal and (iii) where M is chiral ($^C M$) and both enantiomers are present in the crystal. The latter two must be considered as different possible outcomes when crystallising from a racemic solution.¹⁵

If the main component of the solvate is achiral ($^A M$), the situation is almost ideal for limiting the search space in crystal structure prediction calculations, as fewer space groups are

needed, compared to general statistics, to cover the same percentage of observed structures. For example, for CSP calculations of a DMSO solvate of an achiral molecule, the specific $^A M$ /DMSO statistics give almost an 85% confidence that the observed structure will be found in one of only three space groups ($P\bar{1}$, $P2_1/c$ and $C2/c$); eight space groups are needed to cover the same percentage in the overall CSD statistics. By contrast with DMSO, the three most common space groups cover less than 70% of the possibilities for hydrate structures. Space group selection is not as effective at reducing the search space for hydrates.

For ^CM crystallised from an enantiopure solution, $P2_12_12_1$ and $P2_1$ are the most populated space groups for most families of solvates, as in the general database. However, here the observed trend is less helpful than for ^AM: observed space group statistics are slightly less concentrated on the most populated space groups. Whereas in the general CSD statistics 88% of the structures crystallising in the four most popular Sohnke space groups do so in $P2_1$ and $P2_12_12_1$, a slightly smaller percentage (84%) of ^CM solvates is found in these two space groups, and an even lower proportion for certain specific solvates (e.g. only 77% for ^CM DMSO solvates). This suggests that the third and fourth choices of space groups are more important for particular solvate families and the relative importance of the next most popular space groups ($P1$ and $C2$) is dependent on the nature of the solvent (S).

Exploring the possibilities for ^CM crystallising from racemic solutions is computationally the most expensive situation. If we do not know whether enantiomeric resolution will occur (which depends on the relative stability of the enantiopure and racemic crystals),²⁰ one may want to perform calculations in both the most popular Sohnke and centrosymmetric space groups. In most cases, at least the two most populated Sohnke ($P2_1$ and $P2_12_12_1$) and centrosymmetric ($P2_1/c$ and $P\bar{1}$) space groups are needed to cover a reasonable proportion of the likely outcomes.

Finally, we are aware of the small datasets available for some of these subsets of solvates. Therefore, where the number of structures in a particular subset is low, the general solvate distributions (with consideration of chirality) are more reliable, while the space group distributions for ^CM/^AM with a particular solvent molecule may be used when the number of structures is large. As the CSD grows and more solvate crystal structures are reported, the statistical significance of the observed trends for the least populated subsets will improve.

Conclusions

The statistics presented above may serve as a starting point for making an initial selection of the most important space groups for structure generation in crystal structure prediction calculations of solvates. The chirality of the main component and the relevant statistics for the solvent of interest provide appropriate short lists of the most likely space groups. Although the best approach for crystal structure prediction is to consider as many space groups as possible, thus covering as much of phase space as possible, calculations are often limited by the available computing resources. The CSD statistics presented here should help prioritise the most likely space groups in which observed crystal structures will be found. We find that the space group populations for solvate crystal structures differ significantly from the overall CSD statistics. Most importantly, the proportion of crystal structures found in the few most populated space groups is even

higher for solvates than in the CSD as a whole. Therefore, in some cases, a small loss in sampling space is accompanied by an enormous reduction of computational effort required. Thus, in particular cases, careful use of knowledge encapsulated in the CSD can help make a problem, as demanding as structure prediction of solvates, tractable.

Acknowledgements

We would like to thank Drs Neil Feeder, Amy Gillon and László Fábián for fruitful discussions. We thank the Pfizer Institute for Pharmaceutical Materials Science for funding.

References

- 1 B. P. van Eijck and J. Kroon, *J. Comput. Chem.*, 1999, **20**, 799–812.
- 2 R. J. Gdanitz, *Chem. Phys. Lett.*, 1992, **190**, 391–396.
- 3 J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 697–714.
- 4 W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 647–661.
- 5 G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. Scheraga, Y. A. Arnautova, M. U. Schmidt and P. Verwer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2005, **61**, 511–527.
- 6 F. H. Allen, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **B58**, 380–388.
- 7 <http://www.ccdc.cam.ac.uk/products/csd/statistics/>.
- 8 G. M. Day, J. Chisholm, N. Shan, W. D. S. Motherwell and W. Jones, *Cryst. Growth Des.*, 2004, **4**, 1327–1340.
- 9 A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell and W. Jones, *J. Am. Chem. Soc.*, 2006, **128**, 14466–14467.
- 10 B. P. van Eijck, A. L. Speck, W. T. M. Mooij and J. Kroon, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1998, **54**, 291–299.
- 11 I. D. H. Oswald, D. R. Allan, G. M. Day, W. D. S. Motherwell and S. Parsons, *Cryst. Growth Des.*, 2005, **5**, 1055–1071.
- 12 G. M. Day, W. D. S. Motherwell and W. Jones, *Phys. Chem. Chem. Phys.*, 2007, DOI: 10.1039/b612190j.
- 13 B. P. van Eijck and J. Kroon, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 535–542.
- 14 B. P. van Eijck, *J. Comput. Chem.*, 2002, **23**, 456–462.
- 15 F. J. J. Leusen, *Cryst. Growth Des.*, 2003, **3**, 189–192.
- 16 C. H. Gorbitz and H. P. Hersleth, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 526–534.
- 17 A. Nangia and G. R. Desiraju, *Chem. Commun.*, 1999, 605–606.
- 18 I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 389–397.
- 19 E. Pidcock, *Chem. Commun.*, 2005, 3457–3459.
- 20 M. D. Gourlay, J. Kendrick and F. J. J. Leusen, *Cryst. Growth Des.*, 2007, **7**, 56–63.