Prior Likelihoods and Space Group Preferences of Solvates

Jason C Cole, Paul R. Raithby and Robin Taylor*

ABSTRACT: The likelihoods of solvents forming solvates have been estimated by using the recrystallization solvent (RS) data in the Cambridge Structural Database. Although RS data are viewed with caution by some crystallographers, most of the likelihood estimates are shown to have good precision. Strong trends are apparent in the results. For example, high likelihoods are found for aromatic solvents with electron withdrawing substituents, low likelihoods for acyclic aliphatic hydrocarbons. Results for different CSD subsets, such as organic and metalloorganic, are highly correlated. Surprisingly, the likelihood of a solvent to form solvates is almost always higher when the solvent is part of a mixture than when it is pure. This is probably because mixtures are frequently used for substances that are difficult to crystallize. The likelihood of two solvents forming a heterosolvate (i.e., both solvents in the structure) can be well estimated by the product of the likelihoods of the solvents forming normal solvates (i.e., with only one solvent in the structure). It is also shown that (a) the space group preferences of solvates vary significantly with the nature of the cocrystallized solvent, and (b) those of nonsolvates vary significantly with the solvent(s) from which they were crystallized. The possibility exists that the first result is mainly due to the second. There is strong evidence, however, that solvents with inversion centers favor solvate crystallization in centrosymmetric space groups, and solvents with 2-fold rotational symmetry promote crystallization in space groups with 2-fold proper rotational axes. Inclusion of cyclohexane and carbon tetrachloride in a lattice can facilitate crystallization in trigonal and tetragonal space groups, respectively.

1. INTRODUCTION

For many years, the inclusion of solvent[a] in a crystal structure – i.e., solvate or hydrate formation – was little more than a minor irritant to crystallographers, solvent molecules being so often disordered. Since about the turn of the century, however, it has become of real interest. This is largely because of its relevance to pharmaceutical development.[1–5] The inclusion of solvent in the crystal form of an active pharmaceutical ingredient will affect

[a] For the exclusion of doubt, we define a solvent as a compound that is in the liquid state at standard temperature and pressure into which the compound of interest can dissolve.

physicochemical properties such as solubility, dissolution rate, crystal morphology and crystal stability. These in turn may cause pharmacokinetic problems, manufacturing difficulties, unacceptable shelf-life and other unwanted issues. Alternatively, the effects of solvent inclusion are occasionally favorable. The ability to predict solvate and hydrate formation would therefore be very useful,[6–9] and algorithms for this have been published in the last few years.[10–14] Their success rates are typically 80% or higher, which is an encouraging start.

The growth in importance of crystal engineering has also drawn attention to solvates. Rational design of crystal forms requires an understanding of why molecules in crystals arrange themselves as they do. Solvate formation is one aspect of this, and several questions naturally arise, including:

- What interactions do solvents make in crystal structures?[15–22]
- How often do solvates occur?[23]
- Which solvents are most likely to be involved and why?
- Does the inclusion of solvent result in different space group preferences?

It is with the final two of these questions that we are concerned here.

Two papers of particular relevance to this study were published within a year of each other in 1999-2000. Görbitz and Hersleth searched the Cambridge Structural Database (CSD)[24] for solvates and ranked the solvents they contained by their frequencies of occurrence.[23] They established that metalloorganic structures are considerably more likely to include solvents than organic structures. They also found several hundred heterosolvates, i.e., structures containing more than one solvent. Of course, the relative frequencies with which different solvents appear in crystal structures do not tell the whole story, as some are more commonly used as recrystallization solvents than others. This must be taken into account to determine which solvents have the highest propensity to cocrystallize with solutes. Nangia and Desiraju (henceforth ND) defined a measure of this propensity and evaluated it for 20 common organic solvents, restricting their analysis to organic structures.[25] Dimethylformamide and dimethylsulfoxide were found to have the highest propensities. The authors noted that the molecules of both of these solvents can form strong and weak hydrogen-bonds to solute molecules at more than one point. Therefore, they argued, solvent molecules become embedded in solute aggregates during crystal formation.

Recently, a second attempt was made to determine solvate formation propensities. It was part of an interesting study by Cruz-Cabeza, Wright and Bacchi (henceforth CCWB) aimed at establishing whether the propensities are correlated with the entropy penalties associated with solvate formation – as indeed they appeared to be.[26] A plot of the propensity values from this study against those of ND showed a clear correlation but some considerable scatter, raising the question of how precisely solvate formation propensities can be inferred from the CSD.

Investigations into solvate space group preferences have shown that they differ markedly from those of single-component structures.[27–29] Separate space group distributions have been determined for two-component organic solvates in which the main (non-solvent) component is (a) achiral and (b) chiral (the structure being either enantiopure or racemic).[27] This is important information for crystal structure prediction (CSP). CSP is a computationally formidable calculation, so predictions for a given molecular system are usually confined to a selection of the most likely space groups. The results referred to above show that the optimum selection is different for solvates than for unsolvated systems (we are unaware of any antonym for "solvate" so henceforth will use "nonsolvate"). What is less clear is whether the preferred space groups vary according to the nature of the cocrystallized solvents. The 2007 study by Cruz-Cabeza et al. suggested that it might be, but the authors were careful to emphasize that some of their data sets contained only small numbers of structures.[27]

In this paper, we focus on two issues. Firstly, we present new estimates of the propensities of common solvents to form solvates, extending the range of solvents considered in previous work. In contrast to ND and CCWB, our propensities are probabilities, i.e., lie between zero and one. They could serve as prior likelihoods in Bayesian approaches to solvate prediction, where "prior likelihood" refers to the probability of solvate formation from a given solvent before the nature of the solute is taken into account. Hence, we will use this term for our measures of propensity, but shorten it to PL for brevity. We have determined separate PLs for recrystallisation from pure solvents and mixtures, with somewhat surprising results. The uncertainties of the PL estimates have been evaluated. Most importantly, our estimates are derived exclusively from CSD structures for which information is available about the solvents used for crystallization. In this respect, our method differs from those of ND and CC.

Secondly, we re-examine the space group preferences of solvates and, in particular, their dependence on the nature of the cocrystallized solvent. This is not a straightforward issue. If the space group distributions of structures containing solvent X differ from those containing solvent Y, it can be for two reasons. Firstly, the packing behaviors of X and Y may lead to

different space group preferences. For example, Cruz-Cabeza et al. noted that benzene and dioxane are often on inversion centers, and solvates containing those molecules have a particularly pronounced tendency to crystallize in $P2_1/c$ or $P\bar{1}$.[27] Secondly, the chemical structures of solutes crystallised from X may show systematic differences from those crystallised from Y. If this is so, we might expect it to result in different symmetry preferences. These alternative possibilities are investigated.

A key tool in our analysis is the CSD recrystallisation solvent (RS) data field. It is a free format text string that reproduces authors' comments on how they obtained the crystals used for structure determination. Examples are: "*ethyl acetate/hexane*"; "*CH2Cl2*"; "*The material was recrystallised from a mixture of toluene and pentane by solvent layering*". The field seems rarely used and a number of reasons can be suggested. The option to search RS fields in not particularly obvious in the interface of the main CSD search program, ConQuest,[30] although it is present, and also available in the CSD Python API.[31] Because the field is free-format text, a solvent can be referred to in many ways (e.g., *ethanol*, *EtOH*, *abs. alcohol*, etc.). The textual descriptions are occasionally ambiguous (e.g., *xylene* without the *o*-, *m*- or *p*- prefix, perhaps because the crystallographer was unsure*)* and the precise constituents of a mixture are sometimes unclear (most obviously, for the various petroleum ethers). Moreover, the field is unpopulated (empty) for about 82% of CSD entries. Figure 1 shows that almost no RS data were added to the CSD prior to 1996. There was also a sharp fall in annual numbers from 2009 to 2013. A key reason was the increasing number of papers in which crystallography played only a subsidiary role, so that its experimental details were reported in supplementary data (making them much harder to find) or not at all. This and other pressures led the Cambridge Crystallographic Data Centre (CCDC) to launch a new interactive online deposition portal, which summarises key metadata associated with the incoming structure – such as recrystallization solvent – and asks depositors to check and enhance these data at the point at which they are most engaged with the dataset. The steady growth in RS data since 2013 can probably be ascribed to this, and will hopefully continue (the drop at the end of the histogram is an edge effect).

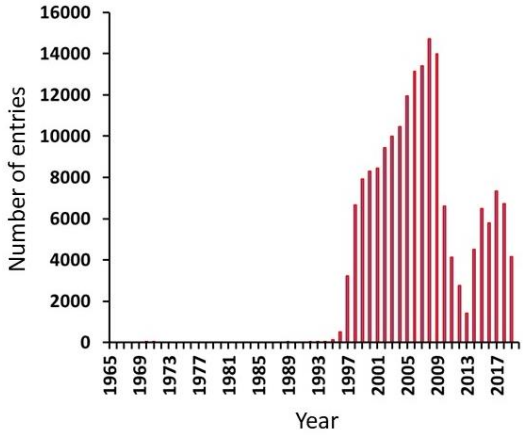

Figure 1. Number of CSD entries with populated recrystallization solvent data fields, by year.

The main reason for the scant use made of the RS field is probably the distrust with which it is viewed by many experienced crystallographers, who are familiar with the shortcomings of day-to-day working practices. For example, solvents in a deposited CIF may refer not to the new structure but to an earlier determination whose CIF was used as a template. Solvents may be mistakenly fitted to residual electron density – *e.g.*, *n*-hexane fitted to a tempting zig-zag of 6 peaks. SQUEEZE[32] may be used to fit residual density with the identity of the solvent being falsely assumed. There may be miscommunications between synthetic chemists and crystallographers that go unobserved because of a perception that the information is unimportant. Finally, much of the data in the CSD can be checked – coordinates must be consistent with published bond lengths, algorithms may pick up misassigned symmetry, etc. Checking RS data is much more difficult and often effectively impossible.

A further aim of this work is therefore to establish the accuracy of RS data, as far as is practicable, and to demonstrate that they can be used effectively.

2. METHODS

2.1. Data Sets. Version 5.41 of the CSD was used (the 2020 release). Entries were rejected if any of the following applied: (a) no atomic coordinates; (b) “unknown solvate” in the compound name (complicates our analysis); (c)”clathrate” in the compound name; (d) entry is a metal-organic framework (i.e., in the CSD MOF subset[33]). We regard structures of types (c) or (d) to be special cases as they are deliberately designed to incorporate guest molecules. When several structures of the same chemical system were available (“refcode families”) only one was retained unless the family contained structures with different space

groups. In that case, one representative of each was kept. Entries with populated RS fields were chosen if possible, the secondary criterion being lowest R-factor.

The resulting data set was divided into subsets, viz., *organic* (restricted to element types H, D; B; C, Si, Ge; N,P, As; O, S; Se; F, Cl, Br, I; noble gases) and *metalloorganic* (contain at least one element not in the preceding list). The organic subset was sometimes divided further into: *achiral*); *chiral* (therefore crystallized in Sohncke space groups); *racemic* (including kryptoracemates[34]). Chirality detection was undertaken using in-house software,[35] a molecule being assigned as chiral only if it had one or more carbon, phosphorus, or sulfur stereocenters and was not a meso-isomer. Chirality due to restricted rotation was not taken into account.

2.2. Parsing of RS Data Fields. Parsing software was written to convert solvent names to standard forms. Occasional errors will have occurred (e.g., because of misspelt names). Solvent names that did not unambiguously define isomeric form, e.g., "xylene" without the *o*-, *m*- or *p*- prefix, were treated as follows. ConQuest searches of the CSD were performed to find solvates containing any of the possible isomeric forms. If the overwhelming majority of the hits corresponded to one of the isomeric forms, this form was assumed when the ambiguous name was encountered. Otherwise, the entry containing the ambiguous name was rejected. For example, "xylene" was rejected but "dichloroethane" was assumed to be 1,2-dichloroethane. Petroleum distillates are common in RS fields, e.g., "light petroleum", "petrol ether". CSD entries with RS fields containing such a name were searched by ConQuest to find alkane solvates. Any alkanes found were equated to the distillate name, e.g. "light petroleum" was converted to "*n*-pentane" and "*n*-hexane" (note that many of the RS fields specify solvent mixtures). Parsing accuracy was estimated by manual inspection of the results obtained for about 500 random entries. There were no misidentified solvents but about 1% of solvent names were not recognised. Table S1 in the Supporting Information lists the occurrence frequencies of the 63 most common organic solvents. The study was confined to these and water.

2.3. Solvate Searches. Structures containing one or more of the most common solvents found in the RS fields were found by exact substructure searching of CSD connectivities (i.e., the chemical connectivities underlying chemical diagrams). This meant that solvates could be found even if the solvent molecules were totally disordered. Structures were not considered solvates if the solvent molecules were all coordinated to metals or metalloids. Entries were

also required to have either the word “solvate” in the compound name or to contain a molecule larger than the solvent molecule(s).

2.4. Estimation of Prior Likelihoods. Consider a given CSD subset and a given solvent. The PL of the solvent is estimated as *nsolvate*/*nsolvent*, where *nsolvent* is the number of entries in the subset whose RS data fields contain the solvent (either alone or with others), and *nsolvate* is the number of those entries in which the crystal structure contains the solvent. In consequence, contributing entries must have populated RS data fields. This reduces their number considerably but has the advantage that every structure used is paired with the solvent(s) used for crystallization. We therefore term the approach the *paired method*. We also tried an *unpaired method* that is more akin to the approaches used by previous investigators; it uses solvate information irrespective of whether or not entries have populated RS fields and extrapolates the solvent distribution in RS fields to the whole CSD. All the PL results below are from the paired method, which we believe to be preferable. Details of the unpaired method and comparison of results from the two methods are in Section S2.

2.5. Statistical Tests. Statistical tests were performed using in-house software (for χ^2 tests), Microsoft Excel® (for 2-way analysis of variance), and the online calculators provided on the Social Science Statistics website[36] (for the remainder). To be acceptable, χ^2 tests on contingency tables require that the expected count in any cell is at least 1 and the proportion of cells with expected counts below 5 does not exceed 20%.[37] When performing such tests on space-group distributions of different solvates or nonsolvates, we iteratively eliminated the lowest-occupancy row (solvent) from the table until these requirements were met. Test results are normally given in parentheses in the text, using the following symbols: r_S: Spearman correlation coefficient; r^2: explained proportion of variance of dependent variable in least-squares regression; *z*: standard normal variate; *dof*: degrees of freedom; *prob*(2T): two-tailed probability of obtaining a test statistic equal to or more extreme than the observed value.

3. ACCURACY OF RECRYSTALLISATION SOLVENT DATA

3.1. Manual Check. A hundred randomly chosen RS fields containing 153 different solvents were checked against the original literature. Most were reported in mainstream synthetic journals where crystallography is usually a side issue. 3 or arguably 4 false negatives were found (solvent incorrectly absent from field), and 3 or 4 false positives (an incorrect solvent in the field). This amounts to an error rate of up to about 5%, albeit based on a very small sample.

3.2. Solvent-Solvate Mismatches. A solvent-solvate mismatch is when a CSD structure contains a solvent that is not named on the entry's populated RS field, suggesting strongly that the field is in error. The combined organic and metalloorganic subsets were searched for mismatches involving any of the solvents included in this study. A large number were found involving water (7610 cases). This is unsurprising because organic solvents are not always anhydrous. Inspection of the mismatches showed they occur most frequently for polar solvents, especially alcohols, as would be expected. Also, if a crystallographer finds a single significant electron density peak a reasonable distance from the molecule being studied, they may assign it to the oxygen of water in the absence of any other ideas. Some of the water mismatches may be due to water vapor sorption.

Unfortunately, leaving water aside still left many mismatches, viz., 3205, which amounted to about 11% of the organic solvents in solvate structures with populated RS fields. Some of these could be ascribed to parsing problems, i.e., when the RS field contained an ambiguous or unrecognised solvent name. The large majority, however, could not. They may be due to several factors, most obviously contamination of solvents and miscommunications between research workers.

Taking 5-year periods from 1995 to 2019, we found that the proportion of entries with populated RS fields that are solvent-solvate mismatches was rather constant during 1995-2010 at about 0.018 – 0.023, but it has fallen to about 0.015 - 0.016 in the last ten years. We can hope that the situation is improving, perhaps because of electronic lab notebooks.

3.3. Assumed Error Rate. The total error rate for organic solvents is much smaller than for water but still substantial. The above results would put it in the range 11 - 16%, depending on how many of the mismatches would have been found had we been able to do manual checking on all solvates (i.e., how much double-counting there would have been). The error rate may be overestimated because it assumes that there are no entries for which the solvent data were correct and the cocrystallized solvents were misidentified by the crystallographers. On the other hand, the situation may be much more serious for nonsolvates. It is likely that the presence of a solvent in a crystal structure reduces the chances of incorrect solvent information being deposited; when the structure is a nonsolvate, there is no convenient reminder of the correct answer. For the purposes of our study, we have therefore assumed an error rate of 35%, i.e., slightly more than double the top end of the error

range deduced above. This pessimism is designed to make our PL standard deviation estimates (Section 4.2) conservative.

4. MODIFICATIONS TO LIKELIHOOD ESTIMATION

In light of the results just described, some modifications were necessary to the method for PL estimation.

4.1. Treatment of Solvent-Solvate Mismatches. When a solvent-solvate mismatch was found, the associated solvent data were changed to the solvent(s) in the crystal structure. This is a partial or complete correction of the error, depending on whether or not other solvents were used in the crystallization (i.e., in a mixture). PLs were then estimated as described in Section 2.4.

4.2. Error Analysis. A standard deviation was estimated for each PL. Two sources of uncertainty are relevant: (a) those arising from deficiencies in the solvent data; (b) sampling errors in the PL estimation.

Uncertainties due to solvent errors were estimated by a simulation consisting of 10,000 cycles of iteration. The procedure within each cycle was:

1. The distribution of solvents in the CSD solvent data for the subset being studied was altered. The number of occurrences of each solvent was arbitrarily increased or decreased by a random amount falling between 0 to 35%. This gave a perturbed solvent distribution for use in step 2.
2. The solvent data associated with a random selection of about 35% of the nonsolvate entries was altered. Each solvent was replaced by something different, chosen from the perturbed solvent distribution (i.e., the probability of a solvent being chosen was proportional to its perturbed number of occurrences).
3. The solvent PLs were calculated from the resulting partially randomized data.

After completion of the cycles, the variance of the 10,000 PL's computed for each solvent were calculated. These measure the effect on PL precision of a 35% solvent error rate.

Each PL is an estimate of a binomial probability, p, from n observations, where n is the number of occurrences of the solvent in the RS data (i.e., *nsolvent*). The Wald estimate of the variance was used, viz., variance = $p(1-p)/n$. The estimate can be unreliable for small sample sizes or probabilities close to zero, where distributions of the p estimate are necessarily

skewed.[38] This is a small problem compared with the gross approximation we have made regarding the error rate of RS data.

The last step is to sum the two variances obtained for each solvent PL and square-root to give the final estimate of standard deviation.

5. RESULTS AND DISCUSSION

5.1. Likelihoods of Solvate Formation. *5.1.1. PLs for Organic and Metalloorganic Subsets*. Tables 1 and 2 show the PL values of solvents with *nsolvent* ≥ 50 for the organic and metalloorganic subsets, respectively (these *nsolvent* values are corrected for solvent-solvate mismatches, Section 4.1). Results for the combined organic plus metalloorganic subsets are in Table S2. Solvents are ordered by decreasing PL.

Table 1. Prior likelihoods of solvents for organic structures

solvent	***nsolvent***	**PL**	**sd**[a]
trifluoroacetic acid[b]	57	0.51	0.07
pyridine	112	0.41	0.06
o-dichlorobenzene	57	0.39	0.07
p-xylene	60	0.37	0.07
1,4-dioxane	244	0.32	0.04
carbon disulfide	133	0.31	0.05
chlorobenzene	134	0.30	0.05
1,2-dichloroethane	211	0.29	0.04
dimethylsulfoxide	722	0.27	0.02
acetic acid	257	0.19	0.03
nitromethane	123	0.19	0.04
benzene	1681	0.19	0.02
dimethylformamide	1327	0.17	0.02
tetrahydrofuran	1445	0.16	0.02
toluene	2542	0.13	0.01
acetonitrile	4181	0.13	0.01
chloroform	7061	0.12	0.01
n-propanol	65	0.12	0.04
n-butanol	102	0.09	0.03
dichloromethane	12742	0.08	0.01

methyl *t*-butyl ether	116	0.08	0.03
methanol	11898	0.07	<0.01
acetone	4590	0.07	0.01
cyclohexane	687	0.06	0.01
isopropanol	1015	0.06	0.01
carbon tetrachloride	182	0.04	0.02
diethyl ether	6042	0.03	<0.01
ethanol	12725	0.03	<0.01
ethyl acetate	9352	0.02	<0.01
di-isopropyl ether	307	0.02	0.01
n-heptane	1051	0.02	<0.01
methyl ethyl ketone	62	0.02	0.02
n-pentane	5384	0.01	<0.01
n-hexane	18117	0.01	<0.01

[a] Estimated standard deviation. [b] This solvent is a special case, see text.

Table 2. Prior likelihoods of solvents for metalloorganic structures

solvent	***nsolvent***	**PL**	**sd**[a]
o-dichlorobenzene	120	0.68	0.06
carbon disulfide	92	0.58	0.06
1,2-dichloroethane	319	0.54	0.04
1,4-dioxane	88	0.51	0.06
chlorobenzene	149	0.48	0.05
fluorobenzene	109	0.43	0.05
acetic acid	62	0.42	0.07
benzene	3053	0.42	0.02
nitromethane	436	0.40	0.03
chloroform	4754	0.36	0.02
dimethylformamide	1437	0.35	0.02
tetrahydrofuran	5474	0.31	0.01
pyridine	390	0.31	0.03
dimethylsulfoxide	728	0.31	0.02
acetonitrile	8067	0.31	0.01

cyclohexane	312	0.29	0.03
o-difluorobenzene	60	0.28	0.06
dichloromethane	24886	0.27	0.01
toluene	6975	0.24	0.01
acetone	3925	0.22	0.01
methanol	7885	0.22	0.01
1,2-dimethoxyethane	325	0.20	0.02
cyclopentane	63	0.19	0.05
methyl *t*-butyl ether	68	0.15	0.05
ethanol	4162	0.12	0.01
isopropanol	500	0.11	0.02
ethyl acetate	764	0.09	0.01
n-octane	102	0.09	0.03
diethyl ether	14687	0.09	<0.01
n-heptane	990	0.07	0.01
di-isopropyl ether	252	0.05	0.01
n-pentane	10011	0.04	<0.01
hexamethyldisiloxane	81	0.04	0.02
n-hexane	21828	0.04	<0.01

[a] Estimated standard deviation.

Two conclusions are immediately obvious. First, the PLs vary greatly between solvents, from close to zero to above 0.5. Second, the uncertainties can be appreciable when *nsolvent* is not large (we often found that this was due more to sampling errors than to the assumed errors in RS data). In particular, the solvents occupying the higher positions in both tables have PLs with high uncertainties, and should therefore be viewed with caution. Moreover, trifluoroacetic acid is a special case because of its low pK_a. All of the solvate structures we used formally contain unionized trifluoroacetic acid but frequently it is accompanied by trifluoroacetate ions, very often in close proximity and with the proton presumably disordered between them. The issue is far less important for acetic acid because of its higher pK_a.

Several trends are apparent. Aromatic solvents with electron withdrawing substituents or ring atoms appear in the upper reaches of the lists. Benzene is significantly higher than toluene in both tables and *p*-xylene is very high in the organic table (it is absent from Table 2 because there are insufficient observations, due to our rejection of entries whose RS data

includes *xylene* without an isomeric qualifier). Aliphatic hydrocarbons are very low in both tables when acyclic, somewhat higher when cyclic. Small, aliphatic solvent molecules with high dipole moments (dimethylsulfoxide, nitromethane, dimethylformamide, acetonitrile) are in the upper parts of both tables. Chlorinated aliphatics are very variable. In both the organic and metalloorganic subsets the order is 1,2-dichloroethane > chloroform > dichloromethane > carbon tetrachloride. Surprisingly, the last of these only had *nsolvent* = 49 for metalloorganics so was excluded from Table 2; it fits the pattern, with PL = 0.18(6) but note the large standard deviation. The relatively low positions of carbon tetrachloride are perhaps because it has no polarized hydrogen atoms.[18,39] Alcohols are in the bottom halves of the tables, with methanol significantly higher than ethanol. Cyclic ethers are high (tetrahydrofuran moderately so, dioxane very high), acyclic ethers low. The only ester, ethyl acetate, is low in both tables and the only ketone, acetone is somewhat higher. The high PLs of carbon disulfide are surprising; we note it is often found in structures of fullerene derivatives.

When the organic and metallorganic subsets are combined, four more solvents satisfy the *nsolvent* ≥ 50 criterion. Their PLs are consistent with the trends just outlined. Specifically, methylcyclohexane and isooctane have low PLs, viz. 0.07(3) and 0.03(2) respectively. Nitrobenzene and benzonitrile are near the top of the table, at 0.56(8) and 0.55(8) respectively. We could find no correlation between the PLs and relative solvent polarities.[40]

As previously noted, metalloorganics are more likely to form solvates than organics;[23] the unweighted average PLs in Tables 1 and 2 are 0.27 and 0.16, respectively. Nevertheless, there is a very good correlation between the organic and metalloorganic PLs (r_S = 0.919, *prob*(2T) < 0.001) except for the outliers labelled in Figure 2. These are dimethylsulfoxide (1) and pyridine (2). The large standard deviation of the organic PL of pyridine may perhaps explain its deviant position but the dimethylsulfoxide figures are more precise, so this solvent may genuinely have unusual behavior. The higher propensity of metalloorganic structures to form solvates is not straightforward to rationalize but it seems likely that they more commonly include awkwardly shaped molecules. A minor factor might be that some solvent molecules in metalloorganic structures make contacts to metal ions that are too long to be classified in CSD as bonds but nevertheless are the reason for the solvent inclusion.

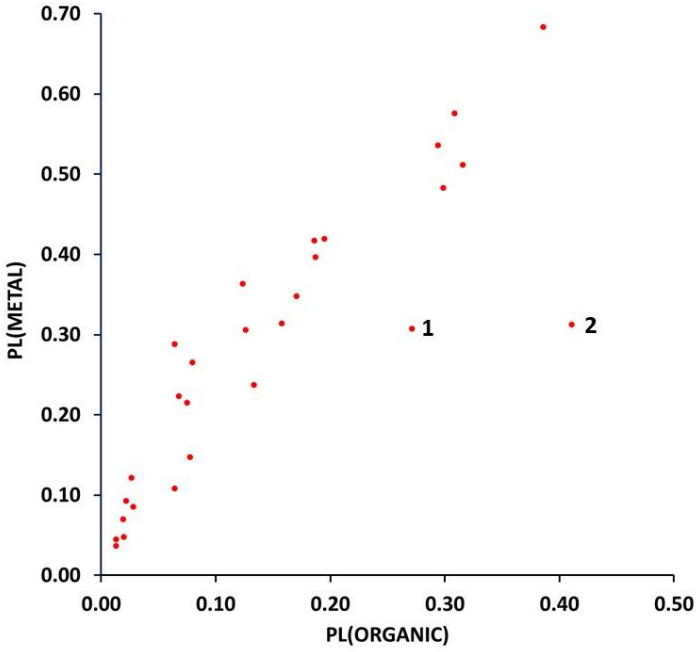

Figure 2. Scatterplot of organic versus metalloorganic PLs.

5.1.2. Average Likelihood. ConQuest searching of entries satisfying our general constraints (3D coordinates present, no clathrates, no unknown solvates, no MOFs) showed the proportion that are solvates to be 0.196 (based on "solvate" in the compound name). In calculating this, we assumed 1.5% of the structures were determined using crystals produced by non-solvent means (from the melt, by sublimation, etc.). In contrast, the average PL over the organic and metalloorganic subsets, weighted by number of solvent occurrences, is only 0.130. This is because a PL measures the probability of a solvent cocrystallizing. The probability of solutes forming solvates is higher because they are often crystallized from mixtures of solvents. Inspection of the RS data fields shows that pure solvents are used for about 50% of crystallizations, the corresponding percentages for 2-, 3- and >3-component mixtures being 45.5%, 4.5% and < 0.2%, respectively. The probability of at least one of the solvents in a 2-component mixture cocrystallizing with the solute can be estimated as $1 - (1 - 0.130)^2 = 0.243$ (i.e., one minus the probability of neither solvent cocrystallizing). The corresponding figure for 3-component mixtures is 0.341. Therefore, the average probability of structures being solvates is 0.50.0.130 + 0.455.0.243 + 0.045.0.341 = 0.189. This is satisfyingly close to the figure from the ConQuest searches and lends credence to the PL estimates.

5.1.3. The PL of Water. Water is absent from Tables 1 and 2 because its PL cannot be estimated with any great confidence. The large number of solvent-solvate mismatches

involving water (Section 3.2) shows beyond doubt that there is a huge tendency to omit water from RS fields when it is, in fact, in the crystallization solvent mixture. For example, in the organic subset entries with RS data there were 4532 hydrates but only 1377 of these were in entries whose RS fields contained "water" or an equivalent such as "aqueous"! This is presumably because of known or unknown use of wet organic solvents, with the former being deemed so obvious (e.g., if the solvent is ethanol) that the research worker does not bother to specify it. Sometimes it may happen by water being taken up by vapor sorption. The cynics among us will also remember that interpreting isolated peaks in residual electron densities as water quickly improves refinements and lowers R-factors.

Solvent-solvate mismatches for water were corrected in the same way as for other solvents (Section 4.1). Unfortunately, it was also necessary to estimate the true number of nonsolvates crystallized from water (pure or in a mixture). This was done by assuming the same error rate for nonsolvates as for solvates (so 4532/1377 for the organic subset, see preceding paragraph).

The resulting PL values are 0.27 for the organics and 0.42 for metalloorganics. We are unable to estimate standard deviations. The values would place water in the top halves of Tables 1 and 2 but not at the very top, and for that reason they may be regarded with scepticism by some. Water is commonly thought of as the solvent most likely to cocrystallize with solutes. Our view is that the ubiquity of hydrates is mainly due to the ubiquity of water in solvent mixtures, and that water does not have the highest inherent preference to cocrystallize with solutes. Cruz-Cabeza et al. also found that water does not have the highest intrinsic propensity to cocrystallize with solutes.[26]

5.1.4. PLs for Organic Achiral, Chiral and Racemic Subsets. Tables S3, S4 and S5 show the PLs of organic solvents in the organic achiral, chiral and racemic subsets. For solvents with *nsolvent* ≥ 50 in all three subsets, the average PLs are 0.09 (achiral), 0.09 (chiral) and 0.08 (racemic), The lower averages compared with those of the organic and metalloorganic subsets (Section 5.1.1) is because several solvents with high PLs in Tables 1 and 2 have *nsolvent* < 50 in at least one of the achiral, chiral, racemic subsets (e.g., pyridine, 1,4-dioxane, carbon disulfide, 1,2-dichloroethane, nitromethane). Pairwise comparisons between the achiral, chiral and racemic subsets, again confined to solvents with *nsolvent* ≥ 50 in both, show good correlations [achiral, chiral: $r_S = 0.833$; achiral, racemic: $r_S = 0.931$; chiral, racemic: $r_S = 0.872$; all with *prob*(2T) < 0.001; Figure S2]. Despite this, the PL of benzene

was appreciably lower in the chiral subset (0.11) than in the achiral (0.20) despite standard deviations of only 0.02. A similar situation was found for tetrahydrofuran (0.07 compared with 0.17). Conversely, the chiral value for isopropanol was higher (0.11 compared with 0.04). A comparison of chiral and racemic subsets shows similar discrepancies. They are possibly due to chemical differences between the chiral and achiral/racemic solutes, e.g., a higher preponderance in the achiral and racemic subsets of molecules that can form stacking interactions with benzene and tetrahydrofuran. Crystallisation of a chiral compound in a chiral space group is perhaps unlikely to be aided by incorporation of a symmetric or planar solvent.

5.1.5. Comparison with Earlier Work. PLs for the organic subset were compared with the propensity measures determined by ND and CCWB, which were also for organic systems. Comparisons were necessarily restricted to the solvents included in their studies.[25,26] Their methods were similar to our unpaired method (Section S2) but the resulting measures were not expressed as probabilities of solvate formation (i.e., in the range 0-1). Nevertheless, their results would be expected to show positive correlations with ours. Visual inspection of the scatterplots (Figure 3a, b) suggests that this is the case, with our results agreeing rather well with those of ND, somewhat less so with those of CCWB (ND: r_S = 0.904, *prob*(2T) < 0.001; CCWB: r_S = 0.704, *prob*(2T) = 0.003). The correlation between ND and CCWB results is very slightly the worst of the three (r_S = 0.692, *prob*(2T) = 0.006).

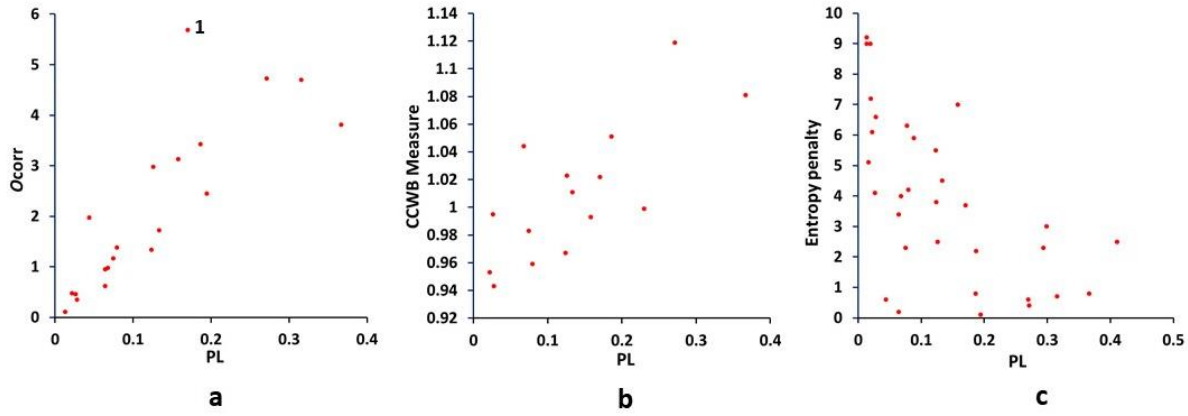

Figure 3. Scatterplots of organic PLs against (a) the ND propensity measure, O_{corr}, (b) the CCWB measure and (c) the entropy penalty data of CCWB. The outlier labelled 1 is dimethylformamide.

ND and CCWB naturally needed estimates of how often each of the solvents they studied is used for crystallization. CCWB obtained this by counting occurrences in the CSD of solvates that have no known unsolvated crystal form (we refer the reader to their Supplementary Information for details). ND used the crystallization-solvent information published in two years' issues of *Acta Crystallographica, Section C*. Both approaches produced solvent counts that were appreciably smaller than the *nsolvent* values in Table 1, which we believe makes our results more reliable. An extreme example is that ND found only 5 uses of dimethylformamide as a crystallization solvent in the journal issues that they perused. It is not hard to imagine that the use of two different years might have produced a different count, and even a difference of 1 would alter the ND propensity value appreciably. Dimethylformamide is the outlying point at the top of Figure 3a (labelled 1) and we suggest this is the reason.

The objective of the CCWB study was to investigate whether the propensity of a solvent to cocrystallize with solutes is related to the entropy penalty of taking the solvent molecules out of the liquid phase. They derived entropy penalty values for 78 solvents and their work suggested that a relationship does exist. We find the same, although the correlation between our probabilities and their entropy penalties is only moderate for the 32 solvents with *nsolvent* ≥ 50 in the organic subset and for which CCWB published entropy penalty values (r_S = -0.643, *prob*(2T) < 0.001). This is perhaps as good as we should expect given that the enthalpies of solvent…solute interactions in the crystalline phase are also relevant to the probability of solvate formation.

5.2. Solvent Mixtures and Heterosolvates. *5.2.1. PLs When Mixtures Used.* Separate PL values were calculated for each solvent, depending on whether recrystallisation was from the pure solvent (PL_p) or a mixture of which it was a component (PL_m). Results were based on the combined organic and metallooorganic subsets set and restricted to the 28 organic solvents for which there were at least 50 entries with RS data for both the pure solvent and the solvent in a mixture.

The results are remarkable (Table S6). For 24 of the 28 solvents, PL_m is higher than PL_p. The median value of PL_m - PL_p is 0.06. The result is highly statistically significant (Wilcoxon test, z = 4.167, *prob*(2T) < 0.001). Some very common solvents have sizeable differences, e.g., $PL_{mixture} - PL_{pure}$ = 0.12 for acetonitrile, 0.08 for dichloromethane. The four solvents for

which $PL_m < PL_p$ are ethyl acetate, *n*-hexane, *n*-pentane and *n*-butanol, but the differences are small and/or not significant.

To explain this result, we perhaps need look no further than paper by Görbitz and Hersleth.[23] They suggested that the use of solvent mixtures should increase the chance of crystallizing high molecular weight solutes. This was based on the reasonable assumption that large molecules often find it difficult to pack efficiently without leaving cavities. The use of a mixture optimises the chance that a solvent is available for filling an inconvenient cavity, with the additional possibility of filling different cavities with different solvents, thus forming a heterosolvate. If this is true, it should mean that researchers more often resort to solvent mixtures when recrystallizing large, awkward molecules that need to solvate, which would explain our result. As a crude test, we determined the mean non-hydrogen counts of the largest component in structures recrystallized from pure solvents and those recrystallized from mixtures. The results were 32.0(1) and 40.9(1), respectively, supporting the explanation.

5.2.2. Heterosolvates. Görbitz and Hersleth suggested that some organic solvents are more likely than others to form heterosolvates, viz.: *n*-hexane, chloroform, dichloromethane, diethyl ether, and toluene and methanol to a smaller extent, when the solute is organic; *n*-pentane, diethyl ether, *n*-hexane, ethanol and methanol when it is metalloorganics.[23] Almost all, they noted, are linear molecules with no branching or aromatic rings. Of course, they had many fewer examples of heterosolvates than are available now so we looked at the matter again, confining ourselves to binary heterosolvates for simplicity.

The most common heterosolvates in the CSD (Table S7) are still in rather good accordance with the results of Görbitz and Hersleth (acetonitrile also appears in high-ranking pairs). These are absolute frequencies of occurrence, however, not taking into account how often each solvent mixture is used. PLs measuring the ability of binary organic solvent mixtures to form heterosolvates were therefore estimated.

Calculations were performed on the combined organic and metalloorganic subsets. The PLs were defined by the usual expression *nsolvate*/*nsolvent*, where *nsolvent* is now the number of entries with RS fields containing the binary solvent mixture under consideration (including when other solvents are in the mixture), and *nsolvate* is the number of those entries in which both of the solvents are in the crystal structure. Results for mixtures with *nsolvent* ≥ 500 are shown in Table 3. The standard deviations in the table are underestimates because

they are based only on the Wald formula (Section 4.2); we did not perform random replacement of mixtures to gauge the effects of RS inaccuracies.

Table 3. Prior likelihoods of binary solvent mixtures forming heterosolvates for the combined organic and metalloorganic subsets

solvent 1	**solvent 2**	***nsolvent***	**PL**[a]
dichloromethane	toluene	626	0.05
acetonitrile	dichloromethane	873	0.04
acetonitrile	methanol	801	0.03
chloroform	methanol	1979	0.03
dichloromethane	methanol	3531	0.02
diethyl ether	tetrahydrofuran	642	0.02
acetonitrile	diethyl ether	3677	0.02
acetone	diethyl ether	1115	0.02
chloroform	diethyl ether	964	0.02
pentane	tetrahydrofuran	1155	0.02
benzene	hexane	1103	0.01
diethyl ether	dimethylformamide	667	0.01
diethyl ether	methanol	1765	0.01
hexane	toluene	1970	0.01
benzene	pentane	586	0.01
dichloromethane	ethanol	1350	0.01
pentane	toluene	997	0.01
dichloromethane	diethyl ether	5659	0.01
chloroform	ethanol	701	0.01
chloroform	pentane	877	0.01
chloroform	hexane	2763	0.01
acetone	ethanol	509	0.01
hexane	tetrahydrofuran	2072	0.01
acetone	hexane	1278	0.00
dichloromethane	pentane	4986	0.00
dichloromethane	hexane	16467	0.00
dichloromethane	heptane	611	0.00

acetone	pentane	566	0.00
diethyl ether	pentane	1225	0.00
ethyl acetate	hexane	4849	0.00
diethyl ether	hexane	2334	0.00
ethyl acetate	pentane	1383	0.00
hexane	pentane	4995	0.00

[a] Standard deviations based on Wald formula all ≤ 0.01; uncertainties due to errors in RS data fields not estimated but likely to be similar.

All the PLs are very low. A reasonable hypothesis is that the PL of a solvent pair X, Y to form a heterosolvate is determined by the product of the individual PLs of X and Y, as calculated in Section 5.1.1 and listed in Table S2. This would be the case if the behaviors of X and Y are independent, the presence of one having no influence on the probability of the other cocrystallizing with the solute. Figure 4 shows the plot of the heterosolvate likelihoods in Table 3 [denoted PL(X,Y)] against the products of the individual likelihoods [denoted PL(X).PL(Y)] taken from Table S2. Also shown is the least-squares regression line, which was constrained to pass through the origin. The line has a gradient of 1.00 and $r^2 = 0.956$, which strongly supports the hypothesis. We conclude that the likelihood of X and Y forming a heterosolvate is almost entirely determined by the product of the individual PLs of X and Y.

Assuming this, the likelihood of *one or both* solvents from a binary mixture cocrystallizing with the solute, PL(X,Y; or X; or Y), should be 1-[1-PL(X)][1-PL(Y)], i.e. one minus the probability of neither cocrystallizing. This also was found to be the case; the least-squares regression line for mixtures with *nsolvent* ≥ 500 was found to have gradient = 1.02, $r^2 = 0.940$.

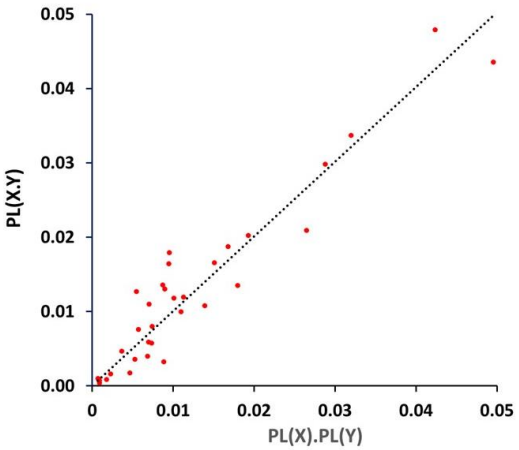

Figure 4. Scatterplot of PL for binary solvent mixtures forming heterosolvates, PL(X,Y), against product of individual solvent likelihoods, PL(X).PL(Y), with least-squares regression line constrained to pass through origin.

5.3. Space Group Distributions of Solvates. Our main focus here was not on whether the space-group preferences of solvates and nonsolvates are different. That has already been clearly established by others; indeed, Grothe et al. showed that multicomponent structures in general have significantly different symmetry preferences to unicomponent structures.[29] Rather, we were interested in whether space group distributions of solvates are significantly influenced by the nature of the solvent they contain. Cruz-Cabeza et al. suggested this was the case but on the basis of a rather small data set, a point they were careful to note.[27]

Space group distributions for the combined organic and metalloorganic subsets are shown in Table 4. It is confined to the ten most common solvate space groups ordered by decreasing popularity. The first line for each solvent shows the distribution for solvates. (Heterosolvates and structures that were both solvates and hydrates were excluded so that the influence on symmetry of individual solvents could be isolated.) The second, italicized line the distribution for nonsolvates crystallized from the solvent (pure or in a mixture). In this context, "nonsolvate" means the structure does not contain the solvent in question. The first number on each row is the number of structures, *n*, which was required to be ≥ 200. It is often lower for the nonsolvates because they were necessarily confined to entries with populated RS fields, whereas this constraint was removed for solvates. The space group occupancies are shown as percentages. Extended tables of solvate space-group distributions for this and other subsets are deposited as Excel® spreadsheets. As Cruz-Cabeza et al. pointed out, the information is useful to guide space-group selection in CSP.[27]

Table 4. Space group occupancies (as percentages) of solvates and nonsolvates.

		space groups[a]									
solvent	**n**[b]	**2**	**14**	**15**	**19**	**4**	**61**	**1**	**5**	**33**	**9**
acetic acid[c]	428	39.7	31.1	7.0	3.7	4.4	1.4	0.7	0.5	0.2	0.2
	269	*28.6*	*38.7*	*8.6*	*4.8*	*5.9*	*1.1*	*0.7*	*1.1*	*0.4*	*1.5*
acetone	5171	33.5	32.3	9.1	4.9	4.1	2.2	1.3	0.9	0.9	1.0
	7891	*21.6*	*36.1*	*7.1*	*9.6*	*6.6*	*4.1*	*1.1*	*1.1*	*1.7*	*1.3*
acetonitrile	14506	36.2	31.9	9.3	3.1	2.8	2.2	0.8	0.5	1.1	1.0
	10168	*24.3*	*38.3*	*8.9*	*5.0*	*3.8*	*3.9*	*0.8*	*0.6*	*1.6*	*1.3*

benzene	8448	37.7	30.8	10.1	2.8	2.4	1.8	0.7	0.6	0.7	0.7
	3298	*23.8*	*40.5*	*8.6*	*6.9*	*3.9*	*3.5*	*0.3*	*0.3*	*1.4*	*1.2*
chloroform	10480	36.6	31.9	7.8	5.5	4.3	2.1	1.1	0.7	0.8	0.7
	9785	*23.6*	*37.8*	*7.5*	*8.6*	*5.9*	*3.8*	*1.1*	*0.9*	*1.2*	*1.1*
cyclohexane	519	31.2	27.4	11.0	3.3	4.0	0.6	1.0	1.3	0.4	0.4
	890	*20.4*	*30.4*	*6.5*	*12.8*	*11.0*	*4.0*	*0.9*	*1.1*	*1.5*	*1.1*
1,2-dichloroethane	852	41.2	32.2	8.8	2.8	2.9	1.9	0.2	0.1	0.7	0.7
	320	*27.5*	*40.0*	*7.5*	*5.0*	*4.1*	*3.8*	*0.3*	*0.0*	*0.9*	*2.5*
dichloromethane	28630	34.8	34.1	8.9	4.2	3.6	2.2	0.8	0.7	1.0	0.9
	31699	*24.3*	*38.7*	*7.2*	*7.6*	*5.2*	*3.9*	*0.8*	*0.6*	*1.6*	*1.2*
diethyl ether	5342	32.9	33.2	9.9	4.4	3.9	2.1	0.7	1.0	1.0	0.7
	20539	*23.8*	*36.9*	*7.4*	*8.2*	*5.8*	*3.9*	*0.7*	*0.6*	*1.6*	*1.2*
1,2-dimethoxyethane	456	35.3	31.4	9.2	5.0	2.9	1.1	0.7	0.2	0.9	0.7
	305	*26.9*	*38.7*	*6.2*	*3.3*	*3.3*	*4.6*	*0.7*	*1.0*	*1.3*	*2.0*
dimethylformamide	4106	40.7	31.1	9.2	2.3	2.3	1.8	1.1	0.4	0.7	1.2
	2219	*25.4*	*39.7*	*9.2*	*3.9*	*2.5*	*3.6*	*0.6*	*0.7*	*1.5*	*1.4*
dimethylsulfoxide	2712	38.5	33.4	7.7	3.2	3.1	1.9	1.4	0.8	0.8	0.7
	1119	*24.4*	*38.5*	*8.6*	*5.9*	*4.4*	*3.8*	*0.9*	*0.9*	*1.7*	*0.8*
1,4-dioxane	771	43.7	30.0	6.7	2.7	2.7	1.3	0.5	0.8	0.3	0.3
	227	*27.3*	*37.4*	*4.8*	*4.0*	*5.7*	*4.0*	*0.9*	*1.8*	*1.3*	*1.3*
ethanol	3971	33.9	31.5	8.5	6.1	5.2	2.3	1.4	1.0	0.7	1.1
	17128	*22.9*	*39.4*	*7.0*	*7.6*	*5.7*	*4.3*	*0.9*	*0.8*	*2.0*	*1.1*
ethyl acetate	1260	30.1	24.3	5.6	11.3	12.5	1.2	2.4	2.3	0.7	0.8
	10316	*18.1*	*33.2*	*4.8*	*16.4*	*11.8*	*3.4*	*1.4*	*1.5*	*1.5*	*1.0*
n-heptane	325	37.5	24.0	11.4	3.4	2.5	2.2	1.5	0.0	0.6	2.8
	2030	*23.2*	*34.2*	*6.7*	*10.6*	*7.9*	*4.3*	*1.7*	*0.9*	*1.4*	*1.4*
n-hexane	4443	40.2	30.7	10.5	2.3	2.4	1.5	0.5	0.6	0.5	0.4
	40424	*24.0*	*37.4*	*6.8*	*9.5*	*6.4*	*3.6*	*1.0*	*0.8*	*1.4*	*0.9*
isopropanol	498	28.7	25.1	4.8	12.0	12.2	2.0	2.2	2.2	0.8	0.4
	1495	*21.9*	*37.1*	*5.9*	*10.4*	*8.0*	*2.8*	*1.9*	*1.5*	*1.1*	*0.8*
methanol	12428	31.7	31.8	7.9	6.6	5.9	2.2	1.5	1.3	1.0	0.6
	18926	*22.3*	*34.5*	*7.0*	*10.8*	*8.0*	*3.5*	*1.5*	*1.1*	*1.4*	*1.3*
nitromethane	687	30.4	34.4	10.9	3.9	2.3	1.7	0.7	0.3	1.0	1.9

	403	*26.8*	*33.5*	*8.4*	*5.0*	*5.0*	*6.0*	*0.5*	*0.7*	*1.0*	*1.5*
n-pentane	2242	35.6	32.4	11.1	3.6	2.2	1.7	0.5	0.6	0.9	0.8
	15454	*24.4*	*39.5*	*6.2*	*8.4*	*5.6*	*3.7*	*0.9*	*0.7*	*1.4*	*1.0*
pyridine	1516	32.1	33.2	10.2	3.6	2.7	1.8	0.5	0.9	1.1	1.3
	357	*25.2*	*34.5*	*10.6*	*4.2*	*5.9*	*3.1*	*0.8*	*1.4*	*0.6*	*1.4*
tetrahydrofuran	10348	31.9	33.6	11.1	3.3	2.6	2.1	0.7	0.5	1.1	1.0
	5254	*25.1*	*38.5*	*8.0*	*5.3*	*3.6*	*4.3*	*0.7*	*0.5*	*1.5*	*1.6*
toluene	11861	39.1	32.3	9.8	2.9	2.1	1.7	0.7	0.4	0.8	0.8
	7817	*24.8*	*39.4*	*7.7*	*5.3*	*3.9*	*4.4*	*0.6*	*0.5*	*1.6*	*0.9*
water	67325	28.3	27.2	9.9	6.2	5.5	2.1	1.5	1.6	1.0	1.0
	5908	*19.0*	*35.3*	*7.2*	*10.3*	*8.0*	*3.4*	*1.7*	*1.3*	*1.7*	*1.0*

[a] Corresponding space group symbols using the most common settings: 2 $P\bar{1}$; 14 $P2_1/c$; 15 $C2/c$; 19 $P2_12_12_1$; 4 $P2_1$; 61 *Pbca*; 1 *P*1; 5 *C*2; 33 $Pna2_1$; 9 *C*c. [b] Number of structures; [c] Here and throughout, first line is solvate distribution, second line nonsolvate.

Differences between solvate and nonsolvate space-group occupancies show very strong trends for the six most common space groups. Occupancy of $P\bar{1}$ is always higher for solvates, as it is for $C2/c$, except for acetic acid, dimethylsulfoxide, isopropanol and pyridine. Nonsolvates have higher occupancy for $P2_1/c$ (except for nitromethane), $P2_12_12_1$ (except for 1,2-dimethoxyethane and isopropanol), $P2_1$ (except for ethyl acetate and isopropanol) and *Pbca* (except for acetic acid). Trends in the various subsets are broadly similar except for enantiopure chiral organics, which must crystallize in Sohnke groups. The most popular space group for these, $P2_12_12_1$, is occupied more often by nonsolvates than solvates for all solvents with $n \geq 100$. Some of the differences in occupancy are extremely large, e.g., 22.5% for benzene, 19.9% for *n*-hexane and 18.1% for diethyl ether. The sample sizes are relatively small, however, so the significance of these extraordinary differences is questionable. $P2_1$ is usually but not always more common for nonsolvates. The reverse is true for $C2$, *P*1, $P2_12_12$ and $C222_1$.

We performed a χ^2 test on the space group distributions of the solvates (using the actual space-group counts, not percentages). Solvents with any value of *n* were included initially and those with the least data were eliminated until (when 27 remained) the conditions for the test were satisfied (Section 2.5). The result was highly significant ($\chi^2 = 4588.5$, *dof* = 234, *prob*(2T) < 0.001). This supports the suggestion by Cruz-Cabeza et al. that the space-group preferences of solvates vary according to the nature of the cocrystallized solvent.[27] Water

makes the dominant contribution to the result because it occurs so much more often in structures than other solvents, but the χ^2 value is still highly significant if it is omitted. Significant results are also obtained on the various CSD subsets, with or without water inclusion.

However, we also established that the nonsolvate space group distributions vary significantly between solvents (χ^2 = 3737.0, *dof* = 234, *prob*(2T) < 0.001). The possibility of errors in the RS fields does not invalidate this result, as they would be expected to reduce differences between the nonsolvate distributions, not increase them; this can be seen by considering the limit where the RS error rate is 100%, making the assignment of entries to solvents completely random. Therefore, the χ^2 result is conservative. Once again, significant results are obtained with the other subsets and without water. The dependence of nonsolvate space group distributions on crystallization solvents is probably because there is a relationship between the chemical nature of a compound and the solvent(s) used to crystallize it, e.g., polar solvents for polar solutes. These differences in chemistry are likely to result in different space group preferences. Another possibility is that solvents may affect the crystallization process,[41] resulting in different symmetry preferences; we consider this far less important,

We are therefore left with a conundrum. The space group preferences of solvates vary significantly with solvent, but why? If systematic differences in solute chemistry are responsible for the different space group preferences of the nonsolvates, then the same reason must apply at least in part to the solvates . Given this, we would expect to observe some similarities in how the solvate and nonsolvate space group preferences vary with solvent. This is indeed the case. Taking $P\bar{1}$ as an example, the percentage occupancies for solvates and nonsolvates (i.e., the numbers in normal and italic font in the relevant column of Table 4) correlate quite strongly (r_S = 0.611, *prob*(2T) = 0.001). Of the next four most common space groups, two show similarly significant correlations ($P2_12_12_1$: r_S = 0.488, *prob*(2T) = 0.013; $P2_1$: r_S = 0.647, *prob*(2T) < 0.001). The r_S values for $P2_1/c$ and $C2/c$ are not significant but they are positive (0.241, 0.170).

Thus, there is no evidence at this point that the *packing* of solvent molecules influences solvate space groups in a solvent-dependent manner. Further statistical testing (analysis of variance on solvate - nonsolvate space-group occupancy differences) failed to resolve this ambiguity (Section S4). We therefore sought other evidence that solvent molecules can have

a direct and noticeable influence on space group symmetries that varies from one solvent to another.

5.4. Solvents on Special Positions. The most likely place to look seemed to be solvent molecules on special positions. It has already been hypothesized that solvent molecules with inversion symmetry (*p*-xylene, benzene, etc.) might produce a bias towards centrosymmetric groups by their tendency to be sited on crystallographic inversion centers.[27] Three types of special positions were investigated: inversion centers, 2-fold proper rotation axes (henceforth, 2-fold axes) and mirror planes. The analysis was limited to organic solvents and was based on the combined organic and metalloorganic subsets, but ignoring solvate structures in which no coordinates were reported for the solvent molecule(s), e.g., because SQUEEZE had been used to fit the electron density.

The first step was to determine the proportions of solvent molecules sited on those positions. This was nontrivial because it was essential to include disordered solvent molecules in the analysis wherever possible; toluene, for example, can only sit on an inversion center if disordered, and frequently does so. Writing code to decide whether a disordered solvent, possibly represented by an incomplete list of disconnected atom sites, is centered on a symmetry element is tricky. Consequently, our results will not be completely accurate, although manual inspection of many examples leads us to believe that we rarely assigned a solvent erroneously to a special position. It was more common for us to make the opposite error. Consequently, the proportions of solvents on special positions reported below are likely to be underestimates. They are expressed as percentages and are termed P_i, P_2 and P_m for percentage of molecules on inversion centers, 2-fold axes and mirror planes, respectively.

The second step was to calculate for each solvent the percentage occupancy of centrosymmetric space groups by solvates (%S_i) and nonsolvates (%N_i) and hence their difference, Δ_i = %S_i - %N_i. The corresponding quantities for space groups with 2-fold axes (Δ_2) and mirror planes (Δ_m) were also calculated. The analysis was restricted to solvents with *nsolvent* ≥ 48 (lower than our normal limit of 50 so we could get the interesting *p*-xylene into the analysis).

It was then straightforward to examine the *P*, Δ relationships. For inversion, the overall correlation between P_i and Δ_i is only moderate but is still highly significant (r_S = 0.610, *prob*(2T) < 0.001; Figure 5a and Table S9). If the solvents investigated are divided into those

that can be sited on a crystallographic inversion center without being disordered (1,4-dioxane, benzene, carbon disulfide, cyclohexane, 1,2-dichloroethane, 1,2-dimethoxyethane, *n*-hexane, *n*-octane and *p*-xylene) and those that cannot, the average Δ_i values of the two sets are 10.4% and 3.8%, respectively. This is a substantial and statistically significant difference (Mann-Whitney z = -2.52, *prob*(2T) = 0.012). The solvents with the largest Δ_i values – in other words, the most effective at promoting centrosymmetric groups – are *p*-xylene, cyclohexane, *n*-hexane and *n*-octane. They are labelled 1-4 on Figure 5a. Points 5 and 6 are 1,4-dioxane and benzene, respectively; as we noted, they have previously been hypothesized to have a role in causing centrosymmetric structures.

A similar situation is found for 2-fold axes (r_S = 0.689, *prob*(2T) < 0.001; Figure 5b and Table S10). Average Δ_2 values for solvents that can and cannot be positioned on a 2-fold axis without being disordered are 4.2 and -0.0%, respectively (Mann-Whitney z = -3.27, *prob*(2T) = 0.001). The two solvents with the highest Δ_2 are carbon tetrachloride and cyclohexane (1 and 2, respectively, on Figure 5b).

In contrast, there is no correlation between P_m and Δ_m (r_S = -0.014, *prob*(2T) = 0.936). This is no surprise because mirror planes are unfavorable to close packing of molecules. The exception is when molecules can occupy the plane, which in the case of solvates ideally means both the solvent and the main component.[42–46]

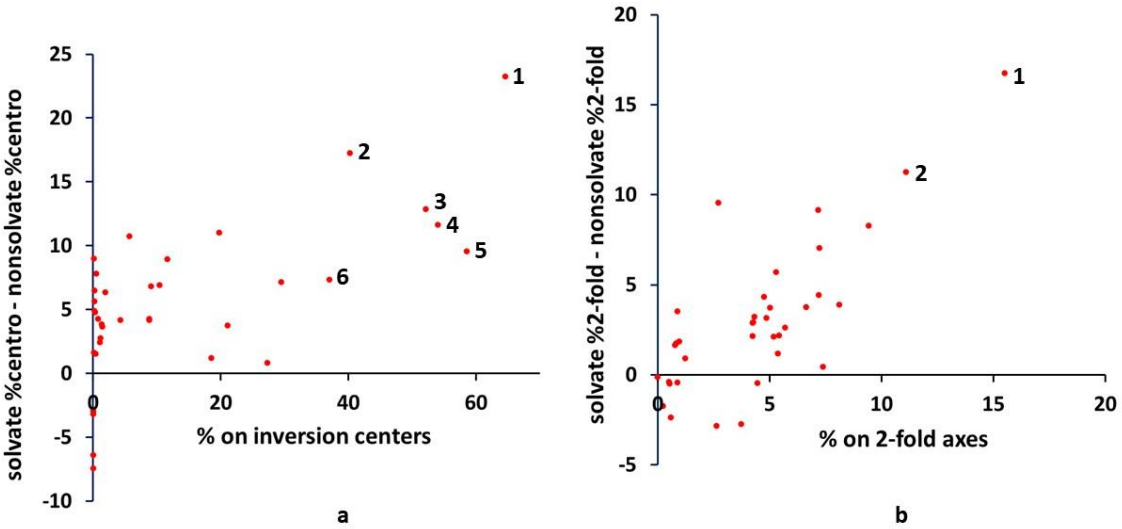

Figure 5. (a) Scatterplot of percentage of solvent molecules on crystallographic inversion centers (P_i) against percentage of centrosymmetric structures for solvates minus percentage for nonsolvates (Δ_i). Labeled points are 1: *p*-xylene; 2: cyclohexane; 3: *n*-hexane; 4: *n*-octane;

5: 1,4-dioxane; 6: benzene. (b) Equivalent plot for 2-fold proper rotation (P_2 against Δ_2). Labeled points are 1: carbon tetrachloride; 2: cyclohexane.

In summary, cocrystallization of solutes with solvents that can be centered on 2-fold axes or inversion centers increases the chances of the space group incorporating those types of symmetry.

5.5. Crystal System Distributions of Solvates. The solvates and nonsolvates were finally categorised by their crystal systems and whether or not their space groups were centrosymmetric, using the combined organic and metalloorganic subsets and excluding heterosolvates. Excel® spreadsheets of the results are available in the Supporting Information. Even a cursory examination reveals two other likely cases of space groups being influenced by solvents on special positions. Cyclohexane and carbon tetrachloride solvates have heightened preferences for trigonal and tetragonal space groups, respectively. In all but 4 of the 38 trigonal cyclohexane solvates with solvent coordinates the molecule is on a 3-fold rotation axis. In 8 of the 15 tetragonal carbon tetrachloride solvates the solvent is on a $\bar{4}$ axis. The structure numbers are small but to us the results seem convincing: these solvents can facilitate crystallization in high-symmetry space groups.

6. CONCLUSIONS

Despite the appreciable error rate of CSD recrystallization solvent data – at least 12.5% and very possibly more – we are confident that our estimates of prior likelihoods are reliable. This is indicated by several factors: their generally low standard deviations; their consistency with the overall frequency of solvates in the CSD; the acceptable correlations between PLs estimated by our preferred unpaired method and those from the alternative paired method (see Figure S1); the excellent correlations between the PLs determined for organic and metalloorganic structures, and between those for achiral and racemic organic structures. The PLs for chiral organics structures correlate a little less well with the other organic subsets but the agreement is still reasonable, and the occurrence of a few discrepancies is unsurprising given the restriction of chiral structures to Sohncke space groups.

Apart from their potential use in Bayesian algorithms for predicting solvate formation, the PLs could be used to guide solvent selection when solvates are unwanted. They may also afford clues about the factors favoring solvate formation. Solvent entropy penalties are surely relevant but the PLs also suggest that the formation of strong intermolecular interactions is a factor. Thus, aromatic solvents with electron withdrawing substituents generally form good

stacking interactions and tend to have high PLs. Nitromethane, dimethylformamide and dimethylsulfoxide have higher than average PLs and can form both hydrogen bonds and strong dipole-dipole interactions. Carbon tetrachloride is below the other chlorinated solvents, probably because it cannot form hydrogen bonds. Alcohols tend to have below average PLs and it is known that their hydroxyl groups can present packing problems.[47] The solvents that form the weakest interactions of all, aliphatic hydrocarbons, tend to have very low PLs, although this must be due in part to entropy factors because cyclohexane and cyclopentane are higher in the table than are their acyclic analogs. Our results are also consistent with the suggestion of Nangia and Desiraju that solvents capable of forming good interactions at multiple points are more likely to form solvates than those that cannot. This would explain why ethers, ketones and esters have lower than average PLs (remembering that ester oxygen atoms rarely accept hydrogen bonds from strong donors[48]).

Water is the most difficult solvent for which to estimate PL values. Identifying how often it is used in crystallizations is extremely difficult because its presence in solvent mixtures is so often unacknowledged. The estimates we have derived are therefore of uncertain reliability. They do not place water at the top of the PL rankings. We suggest it is the most common solvent in crystal structures mainly because it is more often used for crystallizations than organic solvents, often because the latter are sometimes wet.

The work of Grothe et al. made it clear that multicomponent structures in general, including solvates, have distinctly different space group preferences from unicomponent structures.[29] We have now shown that solvates containing different solvents also have different space group preferences. This is almost certainly due in part to systematic chemical differences between solutes crystallized from different solvents. Nevertheless, we have established that the tendency of some solvent molecules to sit on crystallographic inversion centers or 2-fold proper rotation axes is associated with an increase in the percentage of solvates in space groups containing those symmetry elements. Precisely how this comes about in the process of crystal growth is a matter of speculation, but it is likely that molecules such as 1,4-dioxane, benzene and *p*-xylene can fill cavities around inversion centers that would otherwise prevent solutes crystallizing in centrosymmetric space groups. Interestingly, these solvents tend to have higher PLs than close analogs that cannot be sited on an inversion center without being disordered. Thus, *p*-xylene and benzene both have higher PLs than toluene; 1,4-dioxane is higher than tetrahydrofuran; 1,2-dichloroethane is above chloroform and dichloromethane. Perhaps coincidence, perhaps not.

ASSOCIATED CONTENT

Supporting Information

Additional tables (Tables S1 – S10) and figures (Figures S1, S2) on solvent prior likelihoods, description of unpaired method for prior likelihood estimation, and ANOVA analysis of solvate – nonsolvate space group occupancies (PDF). Spreadsheets of space group and crystal system distributions (Excel® .xlsx).

AUTHOR INFORMATION

Corresponding Author

Robin Taylor – *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK;* orcid.org/0000-0002-0391-2609; Email: robin@justmagnolia.co.uk

Authors

Jason C. Cole – *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK;* orcid.org/0000-0002-0291-6317

Paul R. Raithby – *Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK;* orcid.org/0000-0002-2944-0662

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENTS

Matthew Lightfoot and Suzanna Ward of the CCDC are thanked for helpful comments. RT is grateful to the CCDC for an Emeritus Research Fellowship.

REFERENCES

(1) Aaltonen, J.; Allesø, M.; Mirza, S.; Koradia, V.; Gordon, K. C.; Rantanen, J. Solid Form Screening - A Review. *Eur. J. Pharm. Biopharm.* **2009**, *71*, 23–37.

(2) Arora, K. K.; Thakral, S.; Suryanarayanan, R. Instability in Theophylline and Carbamazepine Hydrate Tablets: Cocrystal Formation Due to Release of Lattice Water. *Pharm. Res.* **2013**, *30*, 1779–1789.

(3) Brittain, H. G.; Morris, K. R.; Boerrigter, S. X. M. Structural Aspects of

Solvatomorphic Systems. In *Polymorphism in Pharmaceutical Solids*; Brittain, H. G., Ed.; Informa Healthcare: New York, 2009; pp 233–281.

(4) Griesser, U. J. The Importance of Solvates. In *Polymorphism in the Pharmaceutical Industry*; Hilfiker, R., Ed.; Wiley-VCH: Weinheim, 2006; pp 211–233.

(5) Healy, A. M.; Worku, Z. A.; Kumar, D.; Madi, A. M. Pharmaceutical Solvates, Hydrates and Amorphous Forms: A Special Emphasis on Cocrystals. *Adv. Drug Deliv. Rev.* **2017**, *117*, 25–46.

(6) Boothroyd, S.; Kerridge, A.; Broo, A.; Buttar, D.; Anwar, J. Why Do Some Molecules Form Hydrates or Solvates? *Cryst. Growth Des.* **2018**, *18*, 1903–1908.

(7) Bajpai, A.; Scott, H. S.; Pham, T.; Chen, K.-J.; Space, B.; Lusi, M.; Perry, M. L.; Zaworotko, M. J. Towards an Understanding of the Propensity for Crystalline Hydrate Formation by Molecular Compounds. *IUCrJ* **2016**, *3*, 430–439.

(8) Infantes, L.; Fábián, L.; Motherwell, W. D. S. Organic Crystal Hydrates: What Are the Important Factors for Formation. *CrystEngComm* **2007**, *9*, 65–71.

(9) Cruz-Cabeza, A. J.; Karki, S.; Fábián, L.; Friscić, T.; Day, G. M.; Jones, W. Predicting Stoichiometry and Structure of Solvates. *Chem. Commun.* **2010**, *46*, 2224–2226.

(10) Loschen, C.; Klamt, A. Computational Screening of Drug Solvates. *Pharm. Res.* **2016**, *33*, 2794–2804.

(11) Takieddin, K.; Khimyak, Y. Z.; Fábián, L. Prediction of Hydrate and Solvate Formation Using Statistical Models. *Cryst. Growth Des.* **2016**, *16*, 70–81.

(12) Tilbury, C. J.; Chen, J.; Mattei, A.; Chen, S.; Sheikh, A. Y. Combining Theoretical and Data-Driven Approaches to Predict Drug Substance Hydrate Formation. *Cryst. Growth Des.* **2018**, *18*, 57–67.

(13) Xin, D.; Gonnella, N. C.; He, X.; Horspool, K. Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning. *Cryst. Growth Des.* **2019**, *19*, 1903–1911.

(14) Mohamed, S.; Li, L. From Serendipity to Supramolecular Design: Assessing the Utility of Computed Crystal Form Landscapes in Inferring the Risks of Crystal Hydration in Carboxylic Acids. *CrystEngComm* **2018**, *20*, 6026–6039.

(15) Gillon, A. L.; Feeder, N.; Davey, R. J.; Storey, R. Hydration in Molecular Crystals - A

Cambridge Structural Database Analysis. *Cryst. Growth Des.* **2003**, *3*, 663–673.

(16) Brychczynska, M.; Davey, R. J.; Pidcock, E. A Study of Dimethylsulfoxide Solvates Using the Cambridge Structural Database (CSD). *CrystEngComm* **2012**, *14*, 1479–1484.

(17) Brychczynska, M.; Davey, R. J.; Pidcock, E. A Study of Methanol Solvates Using the Cambridge Structural Database. *New J. Chem.* **2008**, *32*, 1754–1760.

(18) Allen, F. H.; Wood, P. A.; Galek, P. T. A. Role of Chloroform and Dichloromethane Solvent Molecules in Crystal Packing: An Interaction Propensity Study. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2013**, *69*, 379–388.

(19) Zukerman-Schpector, J.; Tiekink, E. R. T. On the Role of DMSO-O(Lone Pair)···π(Arene), DMSO-S(Lone Pair)···π(Arene) and SO···π(Arene) Interactions in the Crystal Structures of Dimethyl Sulphoxide (DMSO) Solvates. *CrystEngComm* **2014**, *16*, 6398–6407.

(20) Infantes, L.; Chisholm, J.; Motherwell, S. Extended Motifs from Water and Chemical Functional Groups in Organic Molecular Crystals. *CrystEngComm* **2003**, *5*, 480–486.

(21) Spiteri, L.; Baisch, U.; Vella-Zarb, L. Correlations and Statistical Analysis of Solvent Molecule Hydrogen Bonding - A Case Study of Dimethyl Sulfoxide (DMSO). *CrystEngComm* **2018**, *20*, 1291–1303.

(22) Alshahateet, S. F.; Bhadbhade, M. M.; Bishop, R.; Scudder, M. L. Different Solvents Yield Alternative Crystal Forms through Aromatic, Halogen Bonding and Hydrogen Bonding Competition. *CrystEngComm* **2015**, *17*, 877–888.

(23) Görbitz, C. H.; Hersleth, H.-P. On the Inclusion of Solvent Molecules in the Crystal Structures of Organic Compounds. *Acta Crystallogr. Sect. B Struct. Sci.* **2000**, *56*, 526–534.

(24) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171–179.

(25) Nangia, A.; Desiraju, G. R. Pseudopolymorphism: Occurrences of Hydrogen Bonding Organic Solvents in Molecular Crystals. *Chem. Commun.* **1999**, 605–606.

(26) Cruz-Cabeza, A. J.; Wright, S. E.; Bacchi, A. On the Entropy Cost of Making

Solvates. *Chem. Commun.* **2020**, *56*, 5127–5130.

(27) Cruz Cabeza, A. J.; Pidcock, E.; Day, G. M.; Motherwell, W. D. S.; Jones, W. Space Group Selection for Crystal Structure Prediction of Solvates. *CrystEngComm* **2007**, *9*, 556–560.

(28) Gavezzotti, A.; Colombo, V.; Lo Presti, L. Facts and Factors in the Formation and Stability of Binary Crystals. *Cryst. Growth Des.* **2016**, *16*, 6095–6104.

(29) Grothe, E.; Meekes, H.; de Gelder, R. Chirality and Stereoisomerism of Organic Multicomponent Crystals in the CSD. *CrystEngComm* **2020**. https://doi.org/10.1039/d0ce00403k.

(30) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr. Sect. B Struct. Sci.* **2002**, *58*, 389–397.

(31) Cambridge Crystallographic Data Centre. CSD Python API https://downloads.ccdc.cam.ac.uk/documentation/API/.

(32) Spek, A. L. PLATON SQUEEZE: A Tool for the Calculation of the Disordered Solvent Contribution to the Calculated Structure Factors. *Acta Crystallogr. Sect. C Struct. Chem.* **2015**, *71*, 9–18.

(33) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29*, 2618–2625.

(34) Fábián, L.; Brock, C. P. A List of Organic Kryptoracemates. *Acta Crystallogr. Sect. B Struct. Sci.* **2010**, *66*, 94–103.

(35) van de Streek, J.; Motherwell, S. New Software for Searching the Cambridge Structural Database for Solvated and Unsolvated Crystal Structures Applied to Hydrates. *CrystEngComm* **2007**, *9*, 55–64.

(36) Social Science Statistics Calculators. https://www.socscistatistics.com/tests/.

(37) Cochran, W. G. Some Methods for Strengthening the Common Chi Square Tests.

Biometrics **1954**, *10*, 417–451.

(38) Brown, L. D.; Cai, T. T.; DasGupta, A. Interval Estimation for a Binomial Proportion. *Stat. Sci.* **2001**, *16*, 101–133.

(39) Desiraju, G. R. The C-H…O Hydrogen Bond in Crystals: What Is It? *Acc. Chem. Res.* **1991**, *24*, 290–296.

(40) Reichardt, C.; Welton, T. *Solvents and Solvent Effects in Organic Chemistry*, 4th ed.; Wiley-VCH: Weinheim, 2011.

(41) Zolotarev, P. N.; Nekrasova, N. A. On the Influence of Solvent Properties on the Structural Characteristics of Molecular Crystal Polymorphs. *Cryst. Growth Des.* **2020**. https://doi.org/10.1021/acs.cgd.0c00753.

(42) Kitaigorodskii, A. I. *Organic Chemical Crystallography*; Consultants Bureau: New York, 1961.

(43) Scaringe, R. P. Electron Crystallography of Organic Molecules; Fryer, J. R., Dorset, D. L., Eds.; Kluwer Academic Publishers: Dordrecht, 1991; pp 85–113.

(44) Filippini, G.; Gavezzotti, A. A Quantitative Analysis of the Relative Importance of Symmetry Operators in Organic Molecular Crystals. *Acta Crystallogr. Sect. B Struct. Sci.* **1992**, *48*, 230–234.

(45) Brock, C. P.; Dunitz, J. D. Towards a Grammar of Crystal Packing. *Chem. Mater.* **1994**, *6*, 1118–1127.

(46) Pidcock, E.; Motherwell, W. D. S.; Cole, J. C. A Database Survey of Molecular and Crystallographic Symmetry. *Acta Crystallogr. Sect. B Struct. Sci.* **2003**, *59*, 634–640.

(47) Brock, C. P.; Duncan, L. L. Anomalous Space-Group Frequencies for Monoalcohols CnHmOH. *Chem. Mater.* **1994**, *6*, 1307–1312.

(48) Lommerse, J. P. M.; Price, S. L.; Taylor, R. Hydrogen Bonding of Carbonyl, Ether, and Ester Oxygen Atoms with Alkanol Hydroxyl Groups. *J. Comput. Chem.* **1997**, *18*, 757–774.