CrystEngComm

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *CrystEngComm*, 2021, **23**, 1555

Organic solvates in the Cambridge Structural Database

Jen E. Werner and Jennifer A. Swift *

Data informatics approaches were applied to the Cambridge Structural Database (CSD) in an effort to discern fundamental trends related to the preparation, occurrence, and general properties of organic solvates. Foremost, the 50 most abundant solvate classes in the CSD were identified through SMILES string matching implemented through CSD Python API, and their relative occurrence rates were compared against data reported 20 years prior. These two sets of data suggest that solvate preparation methods have become less diverse over that time period with an increasing fraction derived from a smaller subset of solvents, though the relative abundance of hetero-solvates containing more than one type of solvent molecule simultaneously increased. A subsequent SMILES string matching facilitated the identification of ~2700 pairs of solvate and solvent-free structures from the top 10 solvate classes. Data from the two related groups showed statistical differences in both the lattice symmetries and packing fractions. Solvates exhibited an inherent bias favoring triclinic lattice symmetry, which is likely related to the larger number of unique molecular components in the asymmetric unit. More surprising was the fact that solvates that do not exhibit disorder statistically had lower packing fractions than their solvent-free analogues. While solvate formation may in fact be a means to achieve phases with higher packing efficiency for some organic molecules, the data indicate this is not a general trend.

Received 1st December 2020,
Accepted 12th January 2021

DOI: 10.1039/d0ce01749c

rsc.li/crystengcomm

Introduction

Crystallization of organic molecules from saturated solutions can lead to either solvated or solvent-free forms.[1–4] The largest class of crystalline solvates are hydrates, owing to the widespread use of water as a solvent and perhaps to a lesser extent the hygroscopicity of many organic molecules.[5] However, many other organic solvents are frequently used in organic synthesis and purification[6] as well as during the solid form screening[7,8] of pharmaceuticals and other commercial materials. This has led to a plethora of crystalline solvate structures in the scientific and patent literatures as well as a growing number of solvated forms in the Cambridge Structural Database (CSD).[9,10]

The role(s) solvent plays in the formation and stability[11] of a solvate can be difficult to pinpoint. In general, strong solute–solvent interactions are thought to play a significant role in predicting solvate formation[12] and these same interactions likely provide some degree of lattice stability that may not be possible in a solvent-free form. Relatedly, it has been shown that as the total polar surface of a molecule increases so does the frequency of hydrate formation.[13,14] Even in the absence of strong solute–solvent interactions, solvate formation has been considered a means to facilitate more efficient space filling, allowing for the generation of more dense phases relative to solvent-free alternatives. This is often used to rationalize why some compounds are prolific solvate formers (*e.g.* gossypol,[15] sulfathiazole,[16] olanzapine[17] axitinib[18] and galunisertib[19]). Even though the number of pharmaceuticals marketed as solvates (other than hydrates) is relatively small[20,21] owing to strict safety[22] and stability requirements, solvates can play an important role in the development process when regarded as precursor phases that can be intentionally desolvated to yield novel solvent-free polymorphs.[23–29] Such process induced transformations become especially relevant when crystal structure prediction methods[30,31] indicate that the lowest energy polymorph has not yet been experimentally realized.

An increasingly popular approach to gain insights into solvate formation and properties is through the statistical analysis of large data sets.[32–38] Though the occurrence rate of solvates (and/or other multicomponent crystals) in the CSD and other industry compilations may differ slightly,[39,40] the CSD remains the largest and most widely accessible source of crystallographic data. The last comprehensive CSD survey of organic solvates was performed by Görbitz and Hersleth in 2000 (October 1998 release)[5] though the number of database entries has grown considerably in the past two decades. Herein we provide an updated analysis of organic solvates in

Georgetown University, Department of Chemistry, Washington, DC 20057-1227, USA. E-mail: jas2@georgetown.edu

the CSD using a structure search method based on simplified molecular input line entry string (SMILES)[41,42] matching which was implemented through the CSD Python application programming interface (API).[43] A similar approach was previously used to analyze hydrates in the CSD.[44] Here, this SMILES string matching method facilitates an updated analysis of the occurrence frequencies of the 50 most common solvents (beyond water), and enables statistical comparisons to be made between solvated and solvent-free forms for the ten most common solvate types.

Curation of the initial organic data set

The first step in the analysis of organic solvates was to curate a working data set of unique structures from the >1 million structures currently in the CSD (version 5.41, February 2020 release).[9] Görbitz and Hersleth[5] had previously shown that the most frequently encountered solvents in organic and metal organic solvates were slightly different, which they noted may be ascribed to the ability of some solvents to act as metal ligands. Since our interest is in organic solvates, we limited the structures considered here to only those with reported 3-D coordinates and the following atoms: H, D, C, N, O, P, S, F, Cl, Br and I.

A second verification step was applied to confirm that each refcode corresponded to a unique polymorph. Duplicate structures were identified with a two-step approach using (1) the "Crystal Packing Similarity" tool in Mercury,[45] and (2) a comparison of unit cell parameters. The Similarity tool compares the ratio of overlapping molecules in two structures for a given packing shell size. A ratio of one means the two structures are identical, with any ratio less than one indicating the two structures are different. However, in the case of solvates, a significant fraction exhibit disorder despite being topologically identical. Analysis with the Similarity tool for these structures is ineffective. Since the disordered entities are no longer a fixed representation of the molecule in the crystal lattice, the Similarity tool will either return a ratio of less than one or fail to converge. Therefore, any refcode pairs with the aforementioned outcomes were subjected to a secondary comparison of the unit cell lengths and angles. Any pair of structures that differed by <1.5% of the largest reduced cell length were treated as identical. When duplicate entries of the same polymorph were identified, the first was retained in the working data set and others removed.

The final curation step was a preemptive validation of the SMILES string associated with each unique structure. In our previous analysis of hydrates,[44] we found that while the vast majority of CSD entries have an entry SMILES string which correctly indicates the component string for each molecule in the crystal, a small fraction of structures had either an incomplete SMILES string or a SMILES string of "none" (~1.6%). All refcodes with SMILES strings were checked for completeness using a text search that verified each solvent molecule in the chemical formula was represented in the entry SMILES string. When an incomplete entry SMILES string was identified, the missing water or organic solvent molecule's component SMILES string was added from a dictionary that linked each solvent to its corresponding formula and compound name(s). The corrected SMILES strings were then used in all subsequent steps. Structures with a SMILES string of "none" were not included. Application of these steps resulted in a final data set consisting of 325 104 unique organic structures.

Search strategy overview

A series of SMILES string searches were applied to the organic data set to identify and categorize different subclasses of solvates as shown in Fig. 1. First, the data set was separated into structures with and without water based on a SMILES string search for water. A second round of SMILES string searches was applied to both the hydrates and the water-free lists to identify entries with a component string corresponding to one or more of the top 50 solvent molecules reported by Görbitz and Hersleth.[5] The use of this list was critical in order to distinguish between solvates and other types of multi-component crystals (*e.g.* cocrystals) in the CSD. Hydrates were separated into "Solvate–Hydrates" (3433) and "Hydrates" (20 850). A similar search of the "Water

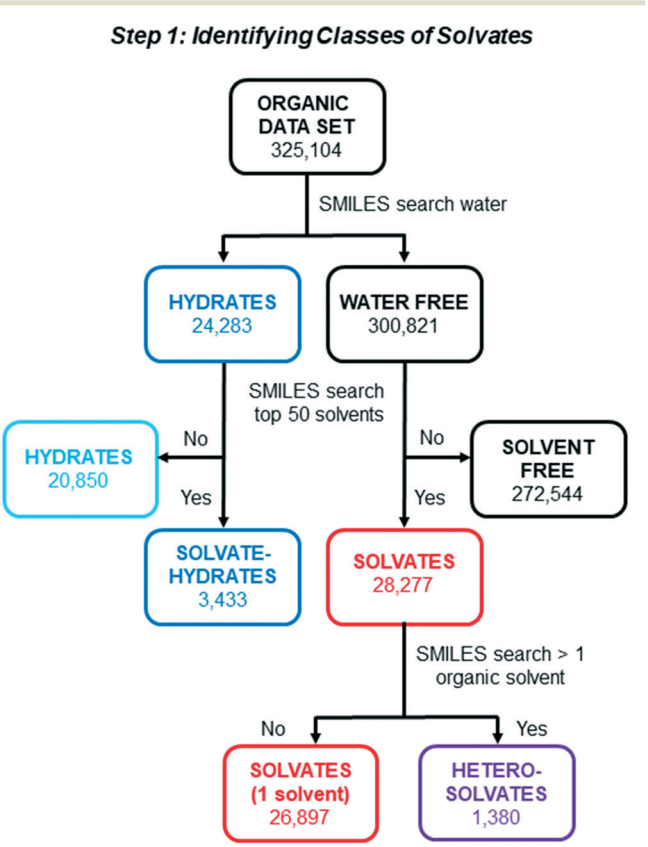

Fig. 1 Flow-chart illustrating steps to identify and sort organic structures in the CSD into five different categories: Solvates, hetero-solvates, solvate–hydrates, hydrates, and solvent-free forms.

Free" list was used to separate the "Solvates" from the larger category of "Solvent Free" structures (note: our use of the term "Solvent Free" does not preclude the possibility that some of the structures with this designation contain solvents other than those in the top 50). The structures in the "Solvates" category were subjected to one final SMILES string search to identify those which contain more than 1 different type of solvent molecule. This final sorting step yielded "Solvates" (26 897) with a single type of solvent, and "Heterosolvates" (1380) with more than one type of non-water solvent component. In total, these search steps identified 31 710 solvated structures, which equates to 9.8%

Table 1 The top 50 organic solvates in the 2020 CSD (V 5.41) ranked according to their frequencies. Numbers in the all solvates column correspond to the total number (and %) of structures containing each solvent. Solvate–hydrate and hetero-solvates list the number of entries (and %) as determined from the total number of solvates for a given solvent molecule. The far right column reproduces numbers from ref. 5

	2020 rank	All solvates (% of total)		Solvate–hydrates (% of that solvent)		Hetero-solvates (% of that solvent)		2000 rank[5]
Methanol	1	5007	(15.79%)	924	(18.5%)	372	(7.4%)	1
Dichloromethane	2	4349	(13.71%)	298	(6.9%)	406	(9.3%)	2
Chloroform	3	4142	(13.06%)	301	(7.3%)	392	(9.5%)	5
Acetonitrile	4	2834	(8.94%)	415	(14.6%)	226	(8.0%)	6
Ethanol	5	1984	(6.26%)	392	(19.8%)	134	(6.8%)	4
Dimethyl sulfoxide	6	1738	(5.48%)	205	(11.8%)	79	(4.5%)	13
Acetone	7	1616	(5.10%)	221	(13.7%)	99	(6.1%)	7
N,*N*-Dimethylformamide	8	1384	(4.36%)	178	(12.9%)	57	(4.1%)	14
Benzene	9	1346	(4.24%)	63	(4.7%)	106	(7.9%)	3
Toluene	10	1171	(3.69%)	58	(5.0%)	97	(8.3%)	8
Ethyl acetate	11	1034	(3.26%)	101	(9.8%)	65	(6.3%)	10
Tetrahydrofuran	12	927	(2.92%)	57	(6.1%)	77	(8.3%)	9
n-Hexane	13	799	(2.52%)	41	(5.1%)	207	(25.9%)	15
Diethyl ether	14	668	(2.11%)	54	(8.1%)	112	(16.8%)	11
Dioxane	15	627	(1.98%)	71	(11.3%)	32	(5.1%)	12
2-Propanol	16	381	(1.20%)	61	(16.0%)	27	(7.1%)	18
Acetic acid	17	341	(1.08%)	51	(15.0%)	7	(2.1%)	16
Pyridine	18	328	(1.03%)	27	(8.2%)	15	(4.6%)	17
1,2-Dichloroethane	19	262	(0.83%)	23	(8.8%)	26	(9.9%)	25
Cyclohexane	20	241	(0.76%)	9	(3.7%)	36	(14.9%)	19
n-Pentane	21	237	(0.75%)	18	(7.6%)	64	(27.0%)	26
p-Xylene	22	233	(0.73%)	10	(4.3%)	18	(7.7%)	20
Carbon disulfide	23	208	(0.66%)	7	(3.4%)	34	(16.3%)	22
Chlorobenzene	24	160	(0.50%)	9	(5.6%)	16	(10.0%)	29
Nitromethane	25	145	(0.46%)	18	(12.4%)	8	(5.5%)	23/24
N,*N*-Dimethylacetamide	26	131	(0.41%)	12	(9.2%)	8	(6.1%)	35/36
1-Propanol	27	108	(0.34%)	21	(19.4%)	11	(10.2%)	27
1,2-Dichlorobenzene	28	104	(0.33%)	2	(1.9%)	16	(15.4%)	35/36
Tetrachloromethane	29	95	(0.30%)	5	(5.3%)	7	(7.4%)	21
n-Heptane	30	87	(0.27%)	12	(13.8%)	30	(34.5%)	37–40
Nitrobenzene	31	79	(0.25%)	10	(12.7%)	8	(10.1%)	23/24
n-Butanol	32	75	(0.24%)	13	(17.3%)	1	(1.3%)	31–33
Formic acid	33	72	(0.23%)	12	(16.7%)	4	(5.6%)	37–40
o-Xylene	34	66	(0.21%)	1	(1.5%)	4	(6.1%)	31–33
m-Xylene	35	64	(0.20%)	1	(1.6%)	12	(18.8%)	30
t-Butanol	36/37	41	(0.13%)	4	(9.8%)	2	(4.9%)	44–46
Ethylene glycol	36/37	41	(0.13%)	9	(22.0%)	1	(2.4%)	42–43
2-Butanol	38	36	(0.11%)	9	(25.0%)	1	(2.8%)	44–46
1,2-Dimethoxyethane	39	32	(0.10%)	6	(18.8%)	1	(3.1%)	28
2-Butanone	40	31	(0.10%)	1	(3.2%)	5	(16.1%)	31–33
Benzonitrile	41	29	(0.09%)	2	(6.9%)	1	(3.4%)	42–43
Propionic acid	42	27	(0.09%)	1	(3.7%)	0	(0%)	47–49
Cyclohexanone	43	26	(0.08%)	5	(19.2%)	1	(3.8%)	37–40
Bromobenzene	44	24	(0.08%)	2	(8.3%)	0	(0%)	37–40
Dibromomethane	45	16	(0.05%)	0	(0%)	0	(0%)	50
Acetophenone	46–48	13	(0.04%)	0	(0%)	0	(0%)	41
Diethyl ketone	46–48	13	(0.04%)	0	(0%)	1	(7.7%)	47–49
Ethylenediamine	46–48	13	(0.04%)	1	(7.7%)	2	(15.4%)	34
1,1,2-Trichloroethane	49	10	(0.03%)	0	(0%)	0	(0%)	44–46
Acetylacetone	50	4	(0.01%)	1	(25.0%)	0	(0%)	47–49
Total		**31 710**[a]		**3433**[a]		**1380**[a]		

[a] Hetero-solvates are counted in each solvent category, making the total listed for each category less than the sum of the numbers in each respective column.

of the working data set. A significantly higher fraction of solvates have one type of solvent molecule (84.8%) than two or more different solvents including water (15.2%).

Top 50 solvates (2020 *vs.* 2000)

Using the search method described, the top 50 solvate formers in the CSD were ranked according to their frequency of occurrence. Table 1 summarizes the distribution for all solvates as well as the subcategories of those with two or more solvents, solvate–hydrates and hetero-solvates. For comparison purposes, the last column reproduces the 2000 rankings previously reported by Görbitz and Hersleth.[5] We note that there are slight differences between our methodology and that used in the previous analysis. The percentages cited in the 2000 statistics include a small number of structures with solvent molecules other than those in the top 50. Since these other solvates constituted only ~5.5% of the total, we assumed their contribution to the current statistical evaluation would be small and did not specifically seek to include them in our search. We also used a more restricted atom list in the creation of our organic data set, which excludes a small number of solvates identified in the previous analysis. Despite the minor differences in how the data sets were curated, the two lists should be directly comparable and accurately reflect any shifting trends over the past two decades.

Overall, the total number of solvate structures in the CSD increased by a factor of ~6, from 5366 in 2000 to 31 710 in 2020, though the increases were unevenly distributed across the different solvent types. The top ten solvate types in 2020 were: (1) methanol, (2) dichloromethane, (3) chloroform, (4) acetonitrile, (5) ethanol, (6) dimethyl sulfoxide (DMSO), (7) acetone, (8) *N*,*N*-dimethylformamide (DMF), (9) benzene, and (10) toluene. These top 10 accounted for just over 80% of all solvates in 2020, up from ~69% in 2000. DMSO and DMF are new to the top 10 list, while ethyl acetate and tetrahydrofuran fell from their former top ten rank in 2000. Expanding to the top 15 solvates, the 2020 and 2000 lists are identical. The top 15 accounted for ~89% of all solvates in 2020, also up from ~85% in 2000. This suggests the typical range of solvents used in the preparation of organic crystals is less diverse than two decades ago.

While the number of CSD entries for solvates in general increased by ~6 times in two decades, the rise of DMSO and DMF to the top 10 list reflects an increase of more than twice that, with relative increases of ~13.5 times and ~12.1 times, respectively. Their entry into the top 10 may in part reflect an increase in the solvents use in other expanding scientific fields.[46,47] Other solvate types which grew at a much faster rate than average include *N*,*N*-dimethylacetamide, chloroform and chlorobenzene with the number of entries increasing by a factor of ~10. In contrast, benzene solvates showed a notable drop in relative rank from #3 to #9, an increase of only ~3 times in the number of reported structures over this same 20 year time period. This may in part be related to a greater awareness of the solvent's toxicological properties.[48,49]

Interestingly, while the data suggests a decrease in the diversity of solvents used over the past two decades, the proportion of solvates with more than one type of solvent molecule increased. Compared to all solvates which increased by ~6 times, solvate–hydrate and hetero-solvate entries increased by ~7.4 times and ~10.5 times, respectively. Highly polar solvents (*e.g.* ethanol, methanol, acetonitrile) were not surprisingly the most frequently encountered in the solvate–hydrates. Trends in top 15 solvate–hydrates generally parallel the total solvates statistics, even though they account for only 10.8% of the total.

On the other hand, hetero-solvates draw from a more diverse combination of solvents. The most common pairs of solvents encountered were methanol–chloroform, methanol–dichloromethane, and dichloromethane–hexane, with 110, 90, and 58 reported structures of each type, respectively. The first two solvent pairs were also among the most common heterosolvates in 2000. In particular, *n*-alkanes (*e.g.* hexane, pentane, heptane) appear to be more likely to crystallize as hetero-solvates than all others in the top 50 list, as evidenced by the significantly higher percentage of *n*-alkane solvates (26.5–47.1%) that contain multiple solvents. In these *n*-alkane heterosolvates, dichloromethane and chloroform are the most common second component.

Solvate stoichiometry

The top ten solvate types were next analyzed on the basis of their stoichiometry. Solvent stoichiometry was determined based on the formula in each refcode entry, since SMILES strings do not distinguish between integral and non-integral numbers of solvent molecules in the asymmetric unit. For each solvate type, those with 1 to 10 solvent molecules were identified in separate lists. Those with >10 solvent molecules were binned into a single collective group. Structures with a non-integral number of solvent molecules were categorized into four groups: hemisolvates (0.5), less than 1, more than 1, and unspecified. Data for all solvate stoichiometries are summarized in Table 2.

In our previous analysis of hydrates[44] we found a strong bias in favor of structures with an integral number of water molecules, with mono- and di-hydrates collectively accounting for 62.7% of all hydrates. A similar bias favoring integral solvent stoichiometries was observed for solvates, however the magnitude of that bias and the general diversity in compositions varied across the solvate types. DMSO and DMF solvates were the most likely to crystallize in ratios of 1 or 2 solvents per host (78.5 and 81.1%). DMSO and DMF solvates were also far less likely than others to have sub-stoichiometric solvent content. To the extent that non-integral solvent compositions could result from partial desolvation of the lattice prior to structure determination, it may be worth noting that these two solvents have significantly higher boiling points than all others in the top

Table 2 Summary of solvate stoichiometries. Both the number of structures and (%) in each category are indicated. The "Hydrates" column reproduces numbers from ref. 44

# Solvent molecules	Hydrates[44]	Methanol	Dichloromethane	Chloroform	Acetonitrile	Ethanol	DMSO	Acetone	DMF	Benzene	Toluene
Integral number of solvent molecules in composition											
1	10 977 (46.3%)	2895 (57.8%)	2183 (50.2%)	2013 (48.6%)	1370 (48.3%)	1203 (60.6%)	937 (53.9%)	850 (52.6%)	782 (56.5%)	536 (39.8%)	515 (44.0%)
2	3893 (16.4%)	713 (14.2%)	600 (13.8%)	796 (19.2%)	491 (17.3%)	222 (11.2%)	427 (24.6%)	255 (15.8%)	341 (24.6%)	148 (11.0%)	159 (13.6%)
3	1092 (4.6%)	132 (2.6%)	119 (2.7%)	181 (4.4%)	155 (5.5%)	44 (2.2%)	72 (4.1%)	56 (3.5%)	43 (3.1%)	54 (4.0%)	44 (3.8%)
4	759 (3.1%)	108 (2.2%)	55 (1.3%)	143 (3.5%)	128 (4.5%)	30 (1.5%)	64 (3.7%)	32 (2.0%)	51 (3.7%)	24 (1.8%)	21 (1.8%)
5	270 (1.1%)	20 (0.4%)	12 (0.3%)	37 (0.9%)	36 (1.3%)	6 (0.3%)	20 (1.2%)	9 (0.6%)	7 (0.5%)	15 (1.1%)	3 (0.3%)
6	273 (1.2%)	29 (0.6%)	11 (0.3%)	34 (0.8%)	44 (1.6%)	5 (0.3%)	13 (0.7%)	8 (0.5%)	23 (1.7%)	7 (0.5%)	7 (0.6%)
7	113 (0.5%)	10 (0.2%)	5 (0.1%)	10 (0.2%)	19 (0.7%)	0 (0%)	9 (0.5%)	4 (0.2%)	5 (0.4%)	3 (0.2%)	3 (0.3%)
8	145 (0.6%)	9 (0.2%)	5 (0.1%)	9 (0.2%)	10 (0.4%)	1 (0.05%)	15 (0.9%)	3 (0.2%)	5 (0.4%)	1 (0.07%)	1 (0.09%)
9	68 (0.3%)	3 (0.06%)	0 (0%)	2 (0.05%)	2 (0.07%)	0 (0%)	4 (0.2%)	1 (0.06%)	1 (0.07%)	2 (0.1%)	0 (0%)
10	81 (0.3%)	0 (0%)	3 (0.07%)	8 (0.2%)	8 (0.3%)	0 (0%)	8 (0.5%)	1 (0.06%)	1 (0.07%)	1 (0.07%)	0 (0%)
>10	467 (1.9%)	7 (0.1%)	10 (0.2%)	7 (0.2%)	10 (0.4%)	3 (0.2%)	21 (1.2%)	1 (0.06%)	1 (0.07%)	4 (0.3%)	1 (0.09%)
Non-integral number of solvent molecules in composition											
0.5	2414 (10.2%)	514 (10.3%)	674 (15.5%)	330 (8.0%)	222 (7.8%)	270 (13.6%)	56 (3.2%)	220 (13.6%)	70 (5.1%)	375 (27.9%)	225 (19.2%)
<1	1242 (5.2%)	263 (5.3%)	391 (9.0%)	244 (5.9%)	121 (4.3%)	127 (6.4%)	26 (1.5%)	84 (5.2%)	21 (1.5%)	84 (6.2%)	73 (6.2%)
>1	1836 (7.7%)	260 (5.2%)	267 (6.1%)	302 (7.3%)	213 (7.5%)	63 (3.2%)	60 (3.5%)	81 (5.0%)	32 (2.3%)	89 (6.6%)	104 (8.9%)
Not specified	68 (0.3%)	44 (0.9%)	14 (0.3%)	26 (0.6%)	5 (0.2%)	10 (0.5%)	6 (0.3%)	11 (0.7%)	1 (0.07%)	3 (0.2%)	15 (1.3%)
TOTAL	**23 698**	**5007**	**4349**	**4142**	**2834**	**1984**	**1738**	**1616**	**1384**	**1346**	**1171**

10. Desolvation seems an unlikely explanation for the particularly low occurrence rates for hemi-solvates compared to other solvate types. In contrast, the fraction of benzene and toluene solvates with 1 : 1 and 1 : 2 host : solvent compositions (50.8 and 57.6%) was significantly lower than all other solvates or hydrates. At least in part this appears to be due to the much higher occurrence rate of 2 : 1 hemi-solvate compositions for these aromatic solvents. *p*-Xylene (rank #22) showed a similarly high occurrence rate for hemi-solvate formation.

Across the different solvate types, the fraction of solvate–hydrates varies although trends largely reflect what might be expected based on a given solvent's miscibility with water. Solvates of alcohols ethanol (19.8%) and methanol (18.5%), followed by acetonitrile (14.6%) and acetone (13.7%) had the largest fraction of solvate–hydrates in the top 10. On the other end of the spectrum, the aromatic solvents benzene (4.6%) and toluene (5.0%) and halogenated solvents dichloromethane (6.8%) and chloroform (7.3%) had the lowest occurrence of solvate–hydrates. Restricting the stoichiometric analysis to only water-free solvates yielded minor changes to the overall statistics, but in general resulted in a modest increase in the fraction of structures with 1 : 1 and 1 : 2 host : solvent compositions.

One of the most noticeable differences between hydrates and other solvates appears to be their ability to access higher stoichiometric ratios, which we define as having 4 or more integral solvent molecules in the composition. In our previous hydrate analysis, we found a natural decrease in the number of structures as the number of water molecules increased, though hydrates with 4 or more water molecules still constituted ~9.2% of the total. When the same analysis was applied to water-free solvates, only acetonitrile (8.8%) and DMSO (8.1%) showed a similar proclivity for adopting compositions with 4 or more solvent molecules. For all other solvates in the top 10, there was a marked decrease in the number of higher solvates, with ethanol (2.2%), dichloromethane (2.2%) and methanol (3.0%) and toluene (3.0%) among the least likely to crystallize with high solvent stoichiometric ratios.

Solvate and solvent-free structure pairs

We next sought to identify from the top 10 solvate lists (21 823 water-free structures) the solvates with known solvent-free crystal forms, with the goal of assessing whether inherent symmetry or density differences exist between the

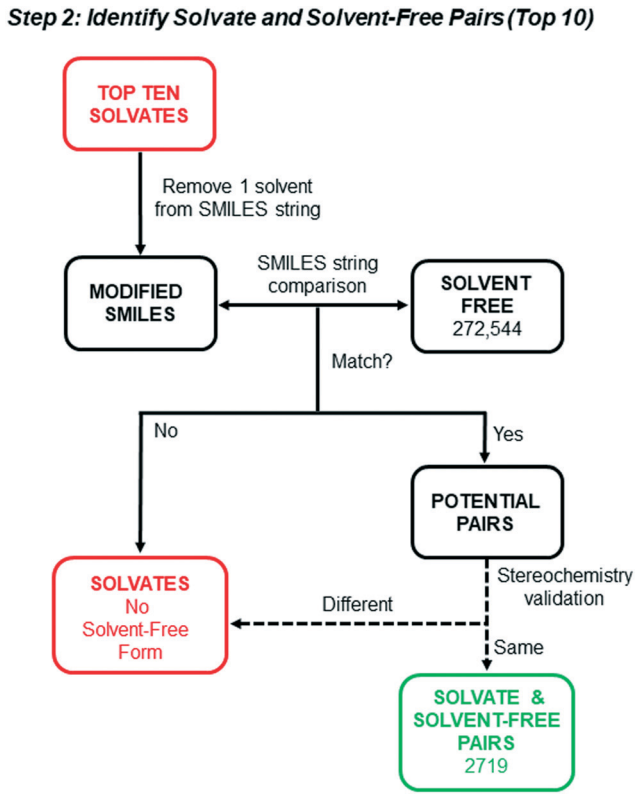

Fig. 2 Flow-chart illustrating steps taken to identify all unique solvate and solvent-free structure pairs. The dashed arrows reflect a hybrid approach consisting of automated chiral center and space group analysis followed by manual comparison.

solvates and non-solvates. Identification of pairs of solvates and their solvent-free counterparts was accomplished *via* SMILES string matching using the steps illustrated schematically in Fig. 2. The entry SMILES string for each refcode contains component strings associated with each unique molecule (solvent and non-solvent) in the crystal. A modified SMILES string was created by removing the solvent component from the entry string. A search of the modified SMILES string against the 272 544 "Solvent Free" structures was then used to generate a list of potential solvate and solvent-free structure pairs.

Due to the inability of SMILES strings to distinguish between stereoisomers, we know this initial list overestimates the number of actual pairs. For example, SMILES string matching will incorrectly match the solvate of a chiral molecule with a solvent-free structure of its enantiomer or the racemate. To identify and remove the false pairs, a sorting method was applied using Python API to identify chiral molecules. An automated search for chiral centers in each non-solvent molecule within the asymmetric unit was carried out. For any structure where chiral centers were identified, the space groups were compared. Chiral molecules crystallizing in centrosymmetric space groups were assumed to be racemic mixtures, whereas chiral molecules in Sohncke space groups where there is one host molecule in the asymmetric unit (ASU) must be homochiral due to a lack of inversion and mirror symmetry.[50] As long as both entries crystallized in one of these space group categories, and if it was a Sohncke group there was one chiral molecule in the ASU of each whose chiral centers matched, they were considered a pair. For any pair of structures where one or both entries had either a non-Sohncke or non-centrosymmetric space group, the molecular contents of the unit cell were manually compared. The same was true if the entries had Sohncke space groups and the chiral molecule appeared more than once in the ASU.

All structures found to have no reported chiral centers had to be investigated manually. For achiral compounds, a quick comparison of the ASU was sufficient. Molecules with asymmetric carbons that eluded the chiral center search or possessing axial chirality went through the same space group assessment to determine whether the unit cells needed to be compared and to what degree. These additional screening steps compensate for all of the stereochemical limitations of the SMILEs string-matching search and led to 2719 "Solvate and Solvent-free Pairs". The pairs reflect ~8.6% of the total number of water-free solvates in the top 10 solvate classes (for comparison purposes, ~6.2% of hydrates had known anhydrate forms). Table 3 shows the breakdown of the number of pairs for each solvate class in the top 10.

Lattice symmetry

All entries in the organic data set, all solvates, all hetero-solvates, and the top 10 solvates with (Table 3) and without known solvent-free forms are sorted according to their seven Bravais lattices in Table 4. The distribution of solvates across the lattices differed significantly from the distribution seen in the organic data set as a whole. Comparatively, "Solvates" showed a large bias favoring low symmetry triclinic lattices (35.6% *vs.* 22.5%). The magnitude of this bias was even greater in the subset of 1380 "Hetero-Solvates" (42.4%). Oddly, "Solvates" also appear to statistically be slightly more likely to adopt higher symmetry tetragonal, trigonal, and hexagonal lattices compared to all structures in the organic data set, though the total fraction of structures with these symmetries is much lower.

Table 3 Number of top 10 solvates (water-free) with a solvent-free counterpart in the CSD. The number of pairs decreases after removal of pairs where one or both structures exhibits disorder and again when the paired structures were determined at different temperatures

Solvent	Total	No disorder	Same temp
Methanol	448	289 (64.5%)	146 (32.6%)
Dichloromethane	272	145 (53.3%)	56 (20.6%)
Chloroform	280	143 (51.1%)	58 (20.7%)
Acetonitrile	288	166 (57.6%)	83 (28.8%)
Ethanol	194	115 (59.3%)	56 (28.9%)
DMSO	344	191 (55.5%)	81 (23.5%)
Acetone	236	160 (67.8%)	83 (35.2%)
DMF	267	187 (70.0%)	96 (36.0%)
Benzene	246	160 (65.0%)	78 (31.7%)
Toluene	144	42 (29.2%)	24 (16.7%)
Total	**2719**	**1598**	**761**

Table 4 Distribution of crystal system symmetry across the organic data set, all solvates, all hetero-solvates and the top 10 (water-free) solvate and solvent-free pairs

	Triclinic	Monoclinic	Orthorhombic	Tetragonal	Trigonal	Hexagonal	Cubic
ORGANIC DATA SET	22.5%	53.2%	21.7%	1.2%	1.0%	0.3%	0.1%
SOLVATES (all)	35.6%	45.5%	14.5%	1.7%	2.1%	0.6%	0.1%
Hetero-solvates (all)	42.4%	41.1%	10.5%	2.2%	3.0%	0.7%	0.1%
TOP 10 PAIRS	34.6%	45.4%	13.4%	1.8%	3.7%	0.9%	0.1%
1 non-solvent molecule	34.1%	46.2%	12.9%	1.8%	4.0%	1.0%	0%
2+ non-solvent molecules	37.9%	40.7%	16.5%	2.0%	2.4%	0%	0.4%
1. Methanol (all)	29.2%	46.3%	20.6%	1.5%	1.8%	0.6%	0.1%
(Pair): solvate	*32.4%*	*42.0%*	*15.4%*	*1.3%*	*8.5%*	*0.4%*	*0%*
: Solvent-free	*21.1%*	*51.6%*	*17.2%*	*0.7%*	*8.3%*	*0.4%*	*0.7%*
2. Dichloromethane (all)	34.8%	46.5%	14.4%	1.6%	2.0%	0.5%	0.2%
(Pair): solvate	*36.0%*	*43.0%*	*15.1%*	*0.7%*	*3.7%*	*1.5%*	*0%*
: Solvent-free	*24.3%*	*49.6%*	*18.8%*	*2.6%*	*2.9%*	*1.5%*	*0.4%*
3. Chloroform (all)	37.9%	44.0%	13.1%	1.7%	2.5%	0.5%	0.3%
(Pair): solvate	*32.9%*	*43.2%*	*16.8%*	*2.1%*	*3.2%*	*1.8%*	*0%*
: Solvent-free	*23.6%*	*57.5%*	*14.3%*	*1.8%*	*2.5%*	*0.4%*	*0%*
4. Acetonitrile (all)	38.4%	44.7%	12.7%	1.9%	1.7%	0.5%	0.1%
(Pair): solvate	*34.0%*	*44.4%*	*11.8%*	*1.0%*	*7.6%*	*0.7%*	*0.3%*
: Solvent-free	*23.6%*	*45.5%*	*16.0%*	*1.0%*	*12.5%*	*0.7%*	*0.7%*
5. Ethanol (all)	33.8%	46.1%	16.4%	1.5%	1.4%	0.7%	0.1%
(Pair): solvate	*41.2%*	*40.7%*	*14.4%*	*0.5%*	*3.1%*	*0%*	*0%*
: Solvent-free	*30.9%*	*47.9%*	*19.1%*	*0%*	*0%*	*0%*	*0%*
6. DMSO (all)	40.7%	46.8%	10.1%	0.8%	1.3%	0.3%	0%
(Pair): solvate	*57.0%*	*26.6%*	*10.1%*	*1.3%*	*5.1%*	*0%*	*0%*
: Solvent-free	*37.5%*	*49.7%*	*9.6%*	*2.0%*	*0%*	*1.2%*	*0%*
7. Acetone (all)	31.8%	47.5%	16.5%	1.7%	1.7%	0.8%	0%
(Pair): solvate	*32.6%*	*46.2%*	*15.7%*	*2.1%*	*3.0%*	*0.4%*	*0%*
: Solvent-free	*26.7%*	*46.2%*	*16.9%*	*2.1%*	*5.1%*	*2.1%*	*0.8%*
8. DMF (all)	46.2%	42.3%	9.7%	0.7%	0.8%	0.3%	0%
(Pair): solvate	*39.2%*	*44.7%*	*6.5%*	*3.7%*	*10.6%*	*1.6%*	*0%*
: Solvent-free	*17.9%*	*56.1%*	*19.5%*	*3.3%*	*2.0%*	*0.4%*	*0.8%*
9. Benzene (all)	40.0%	43.9%	10.3%	1.8%	3.3%	0.6%	0.1%
(Pair): solvate	*44.9%*	*46.4%*	*6.7%*	*0.7%*	*1.1%*	*0%*	*0%*
: Solvent-free	*33.0%*	*47.2%*	*17.2%*	*0.7%*	*0.7%*	*1.1%*	*0%*
10. Toluene (all)	44.3%	40.3%	10.6%	2.0%	2.2%	0.5%	0%
(Pair): solvate	*36.1%*	*40.3%*	*9.0%*	*9.0%*	*4.9%*	*0.7%*	*0%*
: Solvent-free	*25.0%*	*56.2%*	*16.0%*	*0.7%*	*2.1%*	*0%*	*0%*

In our previous analysis of hydrate–anhydrate pairs,[44] the distribution of hydrates across the different lattice types was essentially the same as that of all structures in the working data set. Yet as the number of unique molecules in the lattice increased, the fraction of structures with lower triclinic symmetry also appeared to increase. This was evident from comparisons of hydrate–anhydrate pairs, in which the former by necessity have a higher number of molecules. A trend toward reduced symmetry was also evident when hydrates (with or without anhydrate pairs) were sorted into two categories – those with 1 and 2+ organic components, and the latter were shown to have an even stronger bias toward triclinic lattices.

When a similar analysis was performed on the 2719 solvates and solvent-free structure pairs, the trends were less clear. Whether pairs were considered in the aggregate or treated as separate solvate classes, the fraction of solvates with triclinic structures was consistently higher than the fraction in the solvent-free group. This is consistent with the notion that as the number of unique molecules in the lattice increases, there is a trend toward lower symmetry. However, when the 2719 solvate pairs were sorted into groups with either 1 or 2+ non-solvent molecules, a much more modest change in the distribution across lattice types was observed. In 5 solvate classes (methanol, ethanol, acetone, benzene, and toluene) comparison of structures with 1 to 2+ non-solvent molecules revealed the latter had a higher proportion of triclinic lattices, a trend that paralleled what was seen in the hydrates. In the other 5 solvate classes (dichloromethane, chloroform, acetonitrile, DMSO and DMF), structures with 2+ non-solvent molecules were actually less likely to be triclinic than those with 1 non-solvent molecule. Based on this data, it seems there may be more subtle factors which affect lattice symmetry comparisons across the different solvate classes.

While analysis of compounds which form both solvated and solvent-free forms is an effective means to eliminate some biases, it assumes that the subset of solvates considered is representative of the class of solvates as a whole. In analyzing lattice symmetry specifically, we note that the distribution of lattices across all structures in a given solvate class and the subset with a solvent-free form sometimes differ. This is perhaps most notable in the DMSO and toluene solvates, where the fraction of triclinic structures differs substantially depending on whether all structures in

class or only the subset with known solvent-free forms are considered. Similarly, a disproportionately large fraction of methanol, acetonitrile, and DMF solvates with known solvent-free forms have trigonal lattices relative to all solvates in that class.

Packing fraction

Solvates and solvent-free pairs were additionally compared on the basis of packing fraction. Since the presence of disorder can introduce errors in the calculated packing fraction, only pairs where both structures were ordered were considered. This significantly reduced the number of available pairs to 1598 (Table 3). When a pair was removed from further consideration due to disorder, in the majority of cases (~68%) the solvate exhibited disorder but the solvent-free form was ordered. Of the disordered solvates, over 90% showed disorder in both the solvent and at least one host molecule. In the other 10%, only the solvent was disordered. There were no solvates in which the host molecule(s) were disordered and the solvent ordered. In the other ~32% of pairs that were removed, either both the solvate and solvent-free form exhibited disorder (~17.0%) or the solvate was ordered and the solvent-free form was disordered (~14.0%).

In order to avoid differences due to thermal expansion effects, we further limited our analysis to structure pairs that were determined from data collected at the same temperatures. After this step, 761 pairs remained. The hydrogen atom positions in each structure were normalized, and the packing fraction (PF) was calculated using the packing coefficient algorithm in Mercury. Comparison of the PF of each pair of solvate and solvent-free structures showed that in the vast majority of cases (84%) the difference was 5.0% or less. Each pair was sorted according to the magnitude of the difference between the solvate and solvent-free forms with binning in 0.5% increments. Fig. 3 plots the number of times each solvate (red) or solvent-free (black) structure within a pair has a higher packing fraction. The data suggest that solvates are statistically more likely to have the lower PF, and that the bias appears to be independent of the magnitude of the difference between structure pairs. However, we do not know if this trend would still be apparent if the disordered structures excluded from the analysis had also been considered.

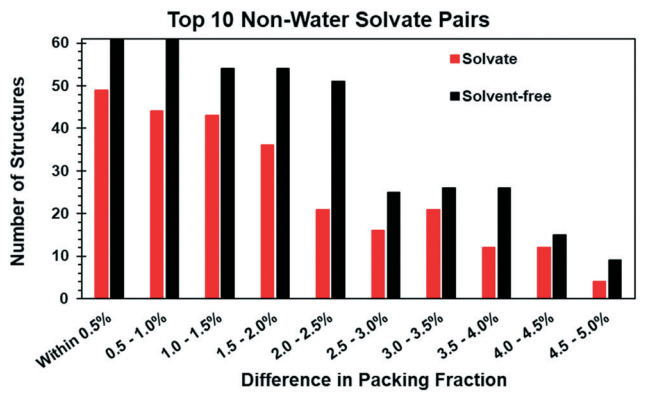

Fig. 3 Comparison of the difference in packing fraction for 640 solvated and solvent-free structure pairs as a function of the difference between the two. Each pair where the solvate has a higher packing fraction are red; pairs where the solvent-free form has a higher packing fraction are black.

That solvates without disorder statistically have lower packing fractions than their solvent-free forms was unexpected, since solvate formation is often rationalized as a means to achieve greater packing efficiency. While this is undoubtedly true in some cases, the data here indicate that statistically speaking that argument does not hold for all solvate classes. Notably, if the top 10 solvate classes are treated individually, the bias is apparent only in about half of the cases. Though the number of pairs in each individual solvate class is low, it is only in ethanol (72.1%), DMSO (69.6%), DMF (68.4%), chloroform (64.9%), acetone (62.9%), and benzene (60.0%) where statistically more solvates have lower packing fractions than their solvent-free forms. The other 5 solvate classes do not show meaningful biases in either direction.

Conclusions

The formation of solvates and the properties they exhibit remain difficult to predict *a priori*. While there will always be a need to analyze individual systems on a case by case basis, we have shown here that data informatics approaches offer an alternative way to gain insight into this class of materials. The CSD is the largest and most widely accessible repository for crystallographic data, and the ability to query it in ways that go beyond the standard Conquest capabilities allows for many new questions to be addressed which can reveal hidden biases and provide evidence that challenges underlying assumptions. We make every effort here to point out potential limitations that exist in both the data and the query methods. This is reflected in steps taken to curate and polish the working data set, and the need to address specific limitations of SMILES string search methods, particularly with respect to stereochemistry.

Comparison of current CSD data against similar data from Görbitz and Hersleth point to some likely shifts over the past two decades in how solvates are generated. It appears that while the range of solvents commonly used in solvate formation has become less diverse, at the same time mixed solvent use has led to a disproportionate increase in reports of hetero-solvates. Relative changes in the growth of individual solvate classes clearly point toward the expanded use of DMSO and DMF, and significantly decreased use of benzene. Organic solvates were also found to adopt a much narrower range of solvent : host stoichiometries compared to organic hydrates.

Direct comparison of the ~2700 pairs of solvates and the solvent-free forms from the top 10 solvate classes indicated differences in both the lattice symmetries and packing densities in the two groups. All solvates were found to have an inherent bias favoring triclinic lattice symmetry, a trend which is especially

magnified in hetero-solvates. Relative increases in the fraction of structures with trigonal lattices were also observed in solvates with known solvent-free forms, though assigning significance to this relative increase may be premature given the number of structure pairs is low. More surprising to us was the fact that solvates without disorder in general, and some solvate classes in particular, showed a bias toward lower packing fractions than their solvent-free analogues. While solvate formation may be a means to achieve phases with higher packing efficiency for some organic molecules, the data indicated this was not an across the board trend.

The two overarching goals of this paper were (1) to assess whether the data indicates practitioner methods for solvate generation have changed over the past 20 years, and (2) to compare some simple metrics based on solvate and solvent-free pairs that might point to hidden structure trends. We hope that the utility of the general approach adopted here can inspire more advanced data mining efforts which address other fundamental questions pertinent to developing a more complete understanding of solvate formation. Toward that end, we stress to practitioners the importance of reporting detailed information on the specific growth conditions employed when new crystal structures are deposited in the CCDC.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank the National Science Foundation (DMR 1609541 and 2004435) and the Henry Luce Foundation (JEW) for financial support.

Notes and references

1 K. R. Morris, Structural aspects of hydrates and solvates, in *Polymorphism in Pharmaceutical Solids*, ed. H. G. Brittain, Marcel Dekker, Inc., New York, 1999, pp. 125–181.
2 S. R. Vippagunta, H. G. Brittain and D. J. W. Grant, Crystalline solids, *Adv. Drug Delivery Rev.*, 2001, **48**(1), 3–26.
3 U. J. Griesser, The importance of solvates, in *Polymorphism in the Pharmaceutical Industry*, ed. R. Hilfiker, Wiley-VCH, Weinheim, 2006, pp. 211–257.
4 A. M. Healy, Z. A. Worku, D. Kumar and A. M. Madi, Pharmaceutical solvates, hydrates and amorphous forms: A special emphasis on cocrystals, *Adv. Drug Delivery Rev.*, 2017, **117**, 25–46.
5 C. H. Görbitz and H.-P. Hersleth, On the inclusion of solvent molecules in the crystal structures of organic compounds, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 526–534.
6 A. C. Society, *Common Solvents Used in Organic Chemistry: Table of Properties*, 2020, https://organicchemistrydata.org/solvents/.
7 J. M. Miller, B. M. Collman, L. R. Greene, D. J. W. Grant and A. C. Blackburn, Identifying the Stable Polymorph Early in the Drug Discover-Development Process, *Pharm. Dev. Technol.*, 2005, **10**, 291–297.
8 J. M. Miller, N. R. Rodriguez-Hornedo, A. C. Blackburn, D. Macikenas and B. M. Collman, Solvent Systems for Crystallization and Polymorph Selection, in *Biotechnology: Pharmaceutical Aspects*, ed. R. T. Borchardt and R. Middaugh, 2007, pp. 53–109.
9 R. Taylor and P. A. Wood, A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts, *Chem. Rev.*, 2019, **119**, 9427–9477.
10 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**(Pt 2), 171–179.
11 A. J. Cruz-Cabeza, S. E. Wright and A. Bacchi, On the entropy cost of making solvates, *Chem. Commun.*, 2020, **56**(38), 5127–5130.
12 S. Boothroyd, A. Kerridge, A. Broo, D. Buttar and J. Anwar, Why Do Some Molecules Form Hydrates or Solvates?, *Cryst. Growth Des.*, 2018, **18**(3), 1903–1908.
13 L. Infantes, J. Chisholm and S. Motherwell, Extended motifs from water and chemical functional groups in organic molecular crystals, *CrystEngComm*, 2003, **5**(85), 480–486.
14 L. Infantes and S. Motherwell, Water clusters in organic molecular crystals, *CrystEngComm*, 2002, **4**(75), 454–461.
15 B. T. Ibragimov, S. A. Talipov and P. M. Zorky, Inclusion complexes of the natural product gossypol, *Supramol. Chem.*, 1994, **3**(2), 147–165.
16 A. L. Bingham, D. S. Hughes, M. B. Hursthouse, R. W. Lancaster, S. Taverner and T. L. Threlfall, Over one hundred solvates of sulfathiazole, *Chem. Commun.*, 2001, 603–604.
17 R. M. Bhardwaj, L. S. Price, S. L. Price, S. M. Reutzel-Edens, G. J. Miller, I. D. H. Oswald, B. F. Johnston and A. J. Florence, Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine, *Cryst. Growth Des.*, 2013, **13**(4), 1602–1617.
18 A. M. Campeta, B. P. Chekal, Y. A. Abramov, P. A. Meenan, M. J. Henson, B. Shi, R. A. Singer and K. R. Horspool, Development of a Targeted Polymorph Screening Approach for a Complex Polymorphic and Highly Solvating API, *J. Pharm. Sci.*, 2010, **99**(9), 3874–3886.
19 R. M. Bhardwaj, J. A. McMahon, J. Nyman, L. S. Price, S. Konar, I. D. H. Oswald, C. R. Pulham, S. L. Price and S. M. Reutzel-Edens, A Prolific Solvate Former, Galunisertib, under the Pressure of Crystal Structure Prediction, Produces Ten Diverse Polymorphs, *J. Am. Chem. Soc.*, 2019, **141**(35), 13887–13897.
20 C. Zhang, K. M. Kersten, J. W. Kampf and A. J. Matzger, Solid-State Insight Into the Action of a Pharmaceutical Solvate: Structural, Thermal, and Dissolution Analysis of Indinavir Sulfate Ethanolate, *J. Pharm. Sci.*, 2018, **107**(10), 2731–2734.
21 B. Bechtloff, S. Nordhoff and J. Ulrich, Pseudopolymorphs in Industrial Use, *Cryst. Res. Technol.*, 2001, **36**(12), 1315–1328.
22 U. S. F. D. Administration, Generally Recognized as Safe (GRAS), (accessed 9/9/2020).
23 A. Bērziņš, A. Trimdale, A. Kons and D. Zvaniņa, On the Formation and Desolvation Mechanism of Organic Molecule Solvates: A Structural Study of Methyl Cholate Solvates, *Cryst. Growth Des.*, 2017, **17**(11), 5712–5724.

24 A. Bērziņš, E. Skarbulis and A. Actiņš, Structural Characterization and Rationalization of Formation, Stability, and Transformations of Benperidol Solvates, *Cryst. Growth Des.*, 2015, **15**(5), 2337–2351.
25 A. Kons, A. Bērziņš, A. Actiņš, T. Rekis, S. van Smaalen and A. Mishnev, Polymorphism of R-Encenicline Hydrochloride: Access to the Highest Number of Structurally Characterized Polymorphs Using Desolvation of Various Solvates, *Cryst. Growth Des.*, 2019, **19**(8), 4765–4773.
26 B. Fours, Y. Cartigny, S. Petit and G. Coquerel, Formation of new polymorphs without any nucleation step. Desolvation of the rimonabant monohydrate: directional crystallisation concomitant to smooth dehydration, *Faraday Discuss.*, 2015, **179**(0), 475–488.
27 J. Mahieux, M. Sanselme and G. Coquerel, Access to Several Polymorphic Forms of (±)-Modafinil by Using Various Solvation–Desolvation Processes, *Cryst. Growth Des.*, 2016, **16**(1), 396–405.
28 D. Martins, M. Sanselme, O. Houssin, V. Dupray, M. N. Petit, D. Pasquier, C. Diolez and G. Coquerel, Physical transformations of the active pharmaceutical ingredient BN83495: enantiotropic and monotropic relationships. Access to several polymorphic forms by using various solvation–desolvation processes, *CrystEngComm*, 2012, **14**(7), 2507–2519.
29 S. Bhattacharya and B. K. Saha, Polymorphism through Desolvation of the Solvates of a van der Waals Host, *Cryst. Growth Des.*, 2013, **13**(2), 606–613.
30 S. L. Price, D. E. Braun and S. M. Reutzel-Edens, Can computed crystal energy landscapes help understand pharmaceutical solids?, *Chem. Commun.*, 2016, **52**(44), 7065–7077.
31 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu and C. R. Groom, Report on the sixth blind test of organic crystal structure prediction methods, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**(4), 439–459.
32 C. J. Tilbury, J. Chen, A. Mattei, S. Chen and A. Y. Sheikh, Combining Theoretical and Data-Driven Approaches To Predict Drug Substance Hydrate Formation, *Cryst. Growth Des.*, 2018, **18**(1), 57–67.
33 K. Takieddin, Y. Z. Khimyak and L. Fábián, Prediction of Hydrate and Solvate Formation Using Statistical Models, *Cryst. Growth Des.*, 2016, **16**(1), 70–81.
34 D. Xin, N. C. Gonnella, X. He and K. Horspool, Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning, *Cryst. Growth Des.*, 2019, **19**(3), 1903–1911.
35 M. Brychczynska, R. J. Davey and E. Pidcock, A study of dimethylsulfoxide solvates using the Cambridge Structural Database (CSD), *CrystEngComm*, 2012, **14**(4), 1479–1484.
36 M. Brychczynska, R. J. Davey and E. Pidcock, A study of methanol solvates using the Cambridge structural database, *New J. Chem.*, 2008, **32**(10), 1754–1760.
37 L. Spiteri, U. Baisch and L. Vella-Zarb, Correlations and statistical analysis of solvent molecule hydrogen bonding – a case study of dimethyl sulfoxide (DMSO), *CrystEngComm*, 2018, **20**(9), 1291–1303.
38 F. H. Allen, P. A. Wood and P. T. A. Galek, 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**(4), 379–388.
39 G. P. Stahly, Diversity in Single- and Multiple-Component Crystals. The Search for and Prevalence of Polymorphs and Cocrystals, *Cryst. Growth Des.*, 2007, **7**(6), 1007–1026.
40 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, Facts and fictions about polymorphism, *Chem. Soc. Rev.*, 2015, **44**(23), 8619–8635.
41 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36.
42 M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys and A. Vaitkus, Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database, *Aust. J. Chem.*, 2018, **10**(1), 23.
43 P. Sanschagrin, Using the CSD Python API for interactive analytics and data mining of the Cambridge Structural Database, *Acta Crystallogr., Sect. A: Found. Adv.*, 2017, **73**(a1), a67.
44 J. E. Werner and J. A. Swift, Data mining the Cambridge Structural Database for hydrate–anhydrate pairs with SMILES strings, *CrystEngComm*, 2020, **22**, 7290–7297.
45 CCDC, Mercury User Guide and Tutorials 2018 CSD Release, No. 800579, 2018.
46 N. C. Santos, J. Figueira-Coelho, J. Martins-Silva and C. Saldanha, Multidisciplinary utilization of dimethyl sulfoxide: pharmacological, cellular, and molecular aspects, *Biochem. Pharmacol.*, 2003, **65**(7), 1035–1041.
47 I. Pastoriza-Santos and L. M. Liz-Marzán, N,N-Dimethylformamide as a Reaction Medium for Metal

Nanoparticle Synthesis, *Adv. Funct. Mater.*, 2009, **19**(5), 679–688.

48 R. Snyder, Recent Developments in the Understanding of Benzene Toxicity and Leukemogenesis, *Drug Chem. Toxicol.*, 2000, **23**(1), 13–25.

49 D. Ross, The Role of Metabolism and Specific Metabolites in Benzene-Induced Toxicity: Evidence and Issues, *J. Toxicol. Environ. Health, Part A*, 2000, **61**(5–6), 357–372.

50 H. D. Flack, Chiral and Achiral Crystal Structures, *Helv. Chim. Acta*, 2003, **86**(4), 905–921.