Supplementary information for:

Space Group Selection for Crystal Structure Prediction of Solvates,

A. J. Cruz-Cabeza, E. Pidcock, G. M. Day, W. D. S. Motherwell and W. Jones

Statistical tests were performed to confirm the significance of differences in space group populations. The Pearson χ^2 test was used for all comparisons.

1) Comparison of RES = 2 populations with the database as a whole

Space group distributions for all structures in the database and structures with RES = 2 (two component crystals) are given in Table 1 (main text). χ^2 was calculated comparing these observed populations (using absolute space group populations, instead of percentage populations). We find a χ^2 of 2384 for the comparison between the two populations which gives a p-value for the significance of differences between the two populations as < 0.001.

The high statistical significance shows that there are real differences in the distributions of crystal structures amongst the space groups between two-component crystal structures and the CSD as a whole. The comparison demonstrates the importance of considering the correct space group populations when making choices for crystal structure prediction calculations, e.g. overall CSD statistics should not be used for predictions on multi-component systems such as solvates.

2) Differences in space group populations by solvent molecule

It appears from the data that the population of space groups of solvates is influenced by the identity of the solvent molecule, so we have performed some statistical analyses of the data, separately for solvates where the main component is chiral and achiral.

i. solvates achiral molecules

Table S1a. Space group frequencies in % for different kind of solvates of achiral molecules.

RES=2 Solvates	N structures	$P2_1/c$ /%	$P\bar{1}$ /%	$P2_12_12_1$ /%	$C2/c$ /%	$P2_1$ /%	$Pbca$ /%	$P1$ /%	$C2$ /%	Rest /%	$P\bar{1}$ +$P2_1/c$ /%
Water	1597	37.0	20.1	5.0	12.4	3.5	4.1	1.1	0.9	15.9	57.1
Methanol	249	39.0	35.3	1.6	5.6	3.6	4.4	0.4	0.0	10.1	74.3
Ethanol	110	41.8	38.2	2.7	5.5	1.8	0.0	1.8	0.9	7.3	80.0
AcOH	55	41.8	43.7	1.8	0.0	1.8	0.0	0.0	0.0	10.9	85.5
EtOEt	38	36.8	15.8	7.9	13.2	7.9	0.0	0.0	0.0	18.4	52.6
AcOEt	51	41.2	37.2	2.0	5.9	5.9	0.0	2.0	0.0	5.8	78.4
Acetone	164	32.9	28.7	4.3	15.3	1.8	3.0	0.6	0.0	13.4	61.6
DMSO	177	31.6	42.4	4.0	10.7	2.3	1.1	0.0	0.0	7.9	74.0
DMF	147	37.4	42.2	2.7	6.1	4.1	0.0	1.4	0.0	6.1	79.6
MeCN	161	42.2	32.3	0.0	8.1	0.6	3.7	0.0	0.0	13.1	74.5
CH_2Cl_2	304	30.9	35.9	3.6	10.8	3.6	3.3	1.3	0.7	9.9	66.8
$CHCl_3$	303	38.3	35.3	3.6	5.3	3.6	2.3	0.7	1.0	9.9	73.6
CCl_4	25	8.0	24.0	12.0	12.0	0.0	4.0	0.0	0.0	40.0	32.0
Dioxane	106	46.2	32.1	0.9	7.6	0.9	0.9	0.0	0.0	11.4	78.3
THF	67	40.3	34.3	4.5	4.5	1.5	0.0	0.0	0.0	14.9	74.6
Benzene	223	29.2	43.0	1.3	10.8	0.4	0.9	0.9	0.9	12.6	72.2
Toluene	96	29.2	44.8	2.1	9.4	1.0	1.0	1.0	0.0	11.5	74.0
Xylene	48	35.4	54.2	4.2	2.1	0.0	0.0	0.0	0.0	4.1	89.6
Hexane	37	21.6	56.8	0.0	10.8	5.4	0.0	0.0	0.0	5.4	78.4
Total solvates	2361	35.6	37.3	2.8	8.3	2.5	1.9	0.7	0.3	10.6	72.9

The sample sizes for many of the solvent molecules are too small for statistical tests of the significance between space group populations. (The general rule of thumb for the χ^2 test is that < 20% of cells in the expected count table should have counts of less than 5 and no cells should have a count of < 1.) To satisfy these sample size criteria for the test, either some of the less common space groups in Table S2a (e.g. *P*1, *C*2) could be merged with the rest of the uncommon space groups, or some of the solvents could be grouped into classes. We did a bit of both.

We first merged the space groups *Pbca*, *P*1 and *C*2 with the rest of the less common space groups, so that we are comparing populations in the 5 most populated space groups and the combined remainder. (Merging *Pbca* was not absolutely necessary for the achiral molecule solvates, but was done to be consistend with what we did for chiral molecules, where a smaller overall sample size meant that only the 5 most common space groups could be treated separately from the rest.) We only want to test that the observed differences are significant and any differences should be apparent in a few of the most popular space groups. Therefore, keeping the 5 most popular space groups separate from the rest should be enough to test the differences between solvent molecules.

We then grouped some of the solvents into chemically similar groups: alcohols (methanol + ethanol); ethyl acetate + diethyl ether; chloromethanes (CH_2Cl_2, $CHCl_3$, CCl_4); benzene + para-xylene. (Toluene was not grouped with benzene and p-xylene because of the suspected importance of the solvent's inversion centre.) Hexane could not be included because of its small sample size. Overall averaged solvate space group populations should be used for solvent molecules with very few known solvate structures, such as hexane, until their number in the database in large enough for more meaningful study.

The final % space group populations for these groupings are given in Table S1b.

Table S1b Space group frequencies in % for the solvates of achiral molecules, for testing of statistical significance.

RES=2 Solvates	**N**	***P*2$_1$/*c* /%**	***P*$\bar{1}$ /%**	***P*2$_1$2$_1$2$_1$ /%**	***C*2/*c* /%**	***P*2$_1$ /%**	**Rest /%**	**constribution to χ^2**
Water	1597	37.0	20.1	5.0	12.4	3.5	22.0	95.0
Methanol + Ethanol	359	39.9	36.2	1.9	5.6	3.0	13.4	18.6
AcOH	55	41.8	43.6	1.8	0.0	1.8	11.0	11.3
EtOEt + AcOEt	89	36.3	28.1	4.5	9.0	6.8	12.3	6.3
Acetone	164	32.9	28.7	4.3	15.2	1.8	17.1	6.1
DMSO	177	31.6	42.4	4.0	10.7	2.3	9.0	17.1
DMF	147	37.4	42.2	2.7	6.1	4.1	7.5	18.4
MeCN	161	42.2	32.3	0.0	8.1	0.6	16.8	11.3
CH_2Cl_2 + $CHCl_3$+ CCl_4	632	33.5	35.1	3.9	8.3	3.5	15.7	9.9
Dioxane	106	46.2	32.1	0.9	7.5	0.9	12.4	8.8
THF	67	40.3	34.3	4.5	4.5	1.5	14.9	3.4
Benzene+ p-Xylene	271	30.2	45.0	1.8	9.3	0.3	13.4	33.9
Toluene	96	29.2	44.8	2.1	9.4	1.0	13.5	10.8

We find an overall $\chi^2 = 251$ against the null hypothesis that the space group populations are the same for all solvent molecules. This shows a high statistical significance (p-value < 0.001 for the 60 degrees of freedom); there are real differences in the space group populations.

The contributions to χ^2 from each solvate type are listed in Table S1b. We found that the greatest contributors to χ^2 are: hydrates > benzene + p-xylene > ethanol + methanol = DMF. These solvates are behaving least like the average behaviour, so it is most important to consider the individual space group populations for these solvent molecules. Those contributing least to the overall χ^2 behave more like the average (e.g. THF), so space group choices in crystal structure prediction for these solvate types could be made based on overall statistics for all solvates.

ii. <u>solvates of chiral molecules</u>

The same criteria for space group merging and grouping of molecules used for the solvates of achiral molecules was applied to solvates of chiral molecules: the 5 most popular space groups were kept separate from the rest (in this case, $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $P2_1$ and $P1$) and the AcOH, THF and hexane solvates could not be treated (samples of 25 or less were too small in this case). The more detailed % populations are given in Table S2a and the final % space group populations, after further space group merging and molecule grouping, are given in Table S2b.

Table S2a. Space group frequencies in % for the solvates of chiral molecules.

RES=2 Solvates	N	$P2_1/c$ /%	$P\bar{1}$ /%	$P2_12_12_1$ /%	$C2/c$ /%	$P2_1$ /%	$Pbca$ /%	$P1$ /%	$C2$ /%	Rest /%	$P2_12_12_1$ +$P2_1$ /%
Water	2195	7.3	4.2	36.0	2.3	26.8	1.1	4.5	6.3	11.5	62.8
Methanol	401	9.0	7.0	34.7	1.5	32.7	1.2	5.5	5.7	2.7	67.4
Ethanol	147	15.0	8.8	28.6	2.7	26.5	0.0	4.1	4.1	10.2	55.1
AcOH	24	8.3	16.7	16.7	0.0	37.5	0.0	4.2	4.2	12.4	54.2
EtOEt	54	11.1	20.4	31.5	3.7	20.4	1.9	3.7	3.7	3.6	51.9
AcOEt	83	10.9	10.9	34.9	2.4	30.1	0.0	6.0	3.6	1.2	65.0
Acetone	146	13.7	14.4	29.4	4.8	19.2	0.7	9.6	2.0	6.2	48.6
DMSO	61	14.7	14.7	26.2	3.3	23.0	0.0	3.3	11.5	3.3	49.2
DMF	38	18.4	26.3	15.8	0.0	23.7	0.0	7.9	2.6	5.3	39.5
MeCN	70	24.3	10.0	20.0	1.4	31.4	1.4	1.4	5.7	4.4	51.4
CH_2Cl_2	151	11.3	19.2	19.2	6.6	27.2	0.0	4.0	4.6	7.9	46.4
$CHCl_3$	131	19.1	21.4	25.2	3.8	16.0	2.3	4.6	4.6	3.0	41.2
CCl_4	14	21.4	14.3	28.6	0.0	14.3	0.0	0.0	0.0	21.4	42.9
Dioxane	26	34.6	30.8	3.8	0.0	23.1	0.0	0.0	3.8	3.9	26.9
THF	19	10.5	15.8	21.0	5.3	15.8	0.0	10.5	0.0	21.1	36.8
Benzene	131	24.4	18.3	15.3	6.9	12.2	0.8	6.9	3.8	11.4	27.5
Toluene	31	12.9	12.9	16.1	3.2	29.0	0.0	12.9	3.2	9.8	45.1
Hexane	25	16.0	32.0	12.0	8.0	16.0	4.0	0.0	4.0	8.0	28.0
Total solvates	1552	14.4	14.1	26.4	3.4	25.1	0.8	5.3	4.6	5.9	51.5

[1.] Para-xylene solvates were not included for chiral molecules, as there were only 7 such solvates in the CSD.

We find an overall $\chi^2 = 366$ against the null hypothesis that the space group populations are the same for all solvent molecules. This shows a high statistical significance (p-value < 0.001 for the 50 degrees of freedom); there are real differences in the space group populations for different solvent molecules.

The contributions to χ^2 from each solvate type are listed in Table S2b. We found that the greatest contributors to χ^2 for solvates of chiral molecules are: hydrates > $CH_nCl_{(4-n)}$ > benzene > dioxane. These solvates are behaving least like the average behaviour, so it is most important to consider the individual space group populations for these solvent molecules.

Table S2b Space group frequencies in % for the solvates of chiral molecules, for testing of statistical significance.

RES=2 Solvates	**N**	***P*2$_1$/*c* /%**	$P\bar{1}$ **/%**	***P*2$_1$2$_1$2$_1$ /%**	***P*2$_1$ /%**	***P*1 /%**	**Rest /%**	**constribution to χ^2**
Water	2195	7.3	4.2	36.0	26.8	4.5	21.2	77.6
Methanol + Ethanol	548	10.6	7.5	33.1	31.0	5.1	12.7	15.1
EtOEt + AcOEt	137	10.9	14.6	33.6	26.3	5.1	9.6	13.4
Acetone	146	13.7	14.4	29.5	19.2	9.6	13.6	20.5
DMSO	61	14.8	14.8	26.2	23.0	3.3	17.9	6.0
DMF	38	18.4	26.3	15.8	23.7	7.9	7.9	24.6
MeCN	70	24.3	10.0	20.0	31.4	1.4	12.9	20.8
CH_2Cl_2 + $CHCl_3$+ CCl_4	296	15.2	19.9	22.3	21.6	4.1	16.8	71.8
Dioxane	26	34.6	30.8	3.8	23.1	0.0	7.7	41.5
Benzene	131	24.4	18.3	15.3	12.2	6.9	22.9	67.0
Toluene	31	12.9	12.9	16.1	29.0	12.9	16.2	8.0