Supporting Information

Rapid Structure Determination of Molecular Solids Using Chemical Shifts Directed by Unambiguous Prior Constraints.

Albert Hofstetter,[†] Martins Balodis,[†] Federico M. Paruzzo,[†] Cory M. Widdifield,[§] Gabriele Stevanato,[†] Arthur C. Pinon,[†] Peter J. Bygrave,[¥] Graeme M. Day[¥] and Lyndon Emsley[†]

[†]Institut des Sciences et Ingéniere Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[§]Oakland University, Department of Chemistry, Mathematics and Science Center, 146 Library Drive, Rochester, MI 48309-4479, United States

[¥]School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

Table of Contents

I. Samples

The powdered samples of anhydrous ampicillin ((2S,5R,6R)-6-([(2R)-2-amino-2-phenylacetyl]amino)-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid, purity > 98.0%) and free base cocaine (Methyl (1R,2R,3S,5S)-3-(benzoyloxy)-8-methyl-8-azabicyclo[3.2.1]octane-2-carboxylate, purity > 98.0%) were purchased from Sigma-Aldrich and Toronto Research Chemicals respectively, while the powdered samples of flutamide (2-Methyl-N-[4-nitro-3-(trifluoromethyl)phenyl]propenamide, purity > 98.0%) and flufenamic acid (2-((3-(Trifluoromethyl)phenyl)amino)benzoic acid, purity > 98.0%) were purchased from Tokyo Chemical Industry. All samples were used without further purification. For all compounds, the reference crystal structures were previously determined by single-crystal XRD. [1-4]

The reference crystal structure of ampicillin (CSD entry: AMCILL) is monoclinic, space group $P2_1$, with unit cell parameters a = 12.40 Å, b = 6.20 Å, c = 12 Å, and 2 molecules in the unit cell.

The reference structure of flutamide, (CSD entry: WEZCOT) contains 4 molecules in the unit cell, and it is orthorhombic, space group $Pna2_1$, with unit cell parameters a = 11.856(2) Å, b = 20.477(3) Å, c = 4.9590(9) Å.

The crystal structure of cocaine, (CSD entry: COCAIN10) contains 2 molecules in the unit cell, it is monoclinic, space group $P2_1$, with unit cell parameters a = 10.130(1) Å, b = 9.866(2) Å, c = 8.445(1) Å.

The flufenamic acid structure (CSD entry: FPAMCA11) is monoclinic, space group P21/c, with unit cell parameters a = 12.523(4) Å, b = 7.868(6) Å, c = 12.874(3) Å and 4 molecules in the unit cell.

II. Solid-state NMR experimental setup

In general, NMR experiments were performed at room temperature on a Bruker 500 wide-bore Avance III and a Bruker 900 US^2 wide-bore Avance Neo NMR spectrometers operating at Larmor frequencies of 500.43 and 900.13 MHz, equipped with H/X/Y 3.2 mm and H/C/N/D 1.3 mm probes.

The 2D ^{1}H-^{13}C dipolar heteronuclear correlation (HETCOR) experiments were performed using a 12.5 kHz MAS frequency for ampicillin, flutamide and cocaine and at 24.0 kHz MAS frequency for flufenamic acid. In all experiments, SPINAL-64 was used for heteronuclear decoupling during t2 and eDUMBO-1$_{22}$ for homonuclear decoupling in the indirect dimension. 16 and 128 transients with 256 t1 increments were acquired for ampicillin, 64 transients and 256 t1 increments for flutamide, 4 transients with 64 t1 increments for flufenamic acid and 16 transients with 256 t1 increments for cocaine.

For the assignment of ampicillin, additional NMR experiments were required. These experiments were performed on either a standard-bore Bruker 700 Avance III or a wide-bore Bruker 500 Avance III operating at Larmor frequencies of 700.04 MHz and 500.16 MHz, respectively. Experiments on the 700 used a 3.2 mm HCN probe, while those on the 500 used a 4.0 mm HX probe. Recycle delays were between 1.0 and 1.3 s for all experiments outlined below.

The 11.7 T 2D ^{13}C-^{13}C refocused Incredible Natural Abundance Double Quantum Transfer Experiment (INADEQUATE) was performed using a 13.0 kHz MAS frequency at a temperature of 295 K. Prior to the indirect evolution period, cross-polarization (CP) from the ^{1}H nuclei was carried out (contact time of 2.5 ms). SPINAL-64 heteronuclear decoupling (100 kHz nutation frequency) was used during both evolution dimensions. 1760 transients with 128 t_1 increments were used. Each τ delay during the indirect dimension evolution was set to 3.84 ms, the length of the z-filter was 1.0 ms, and the recycle delay was 1.0 s.

The 16.4 T ^{1}H-^{15}N CP-HETCOR NMR experiment was carried out at T = 265 K using a 15 kHz MAS rotation frequency, while a ^{15}N magic-angle-turning (MAT) experiment was performed at T = 266 K and a 1.90 MAS rotation frequency. For the ^{1}H-^{15}N HETCOR experiment, SPINAL-64 heteronuclear decoupling was used during the t_2 dimension (83 kHz nutation frequency), and eDUMBO-1$_{22}$ was used for homonuclear decoupling in the indirect dimension (the scaling factor was set to 0.564). Prior to the indirect evolution period, CP from the ^{1}H nuclei was done (contact time = 300 μs). 1440 transients with 64 t_1 increments were used. For the ^{15}N MAT experiment, SPINAL-64 heteronuclear decoupling was used during both the t_1 and t_2 dimensions (100 kHz nutation frequency). Prior to the indirect evolution period, CP from the ^{1}H nuclei was done (contact time = 5.5 ms), with 1024 transients being acquired and averaged per t_1 increment, and with 125 t_1 increments being used. The raw data is available in the Supplementary Material.

The ^{1}H and ^{13}C chemical shifts were referenced indirectly to tetramethylsilane using the methyl signals of L-alanine at 1.3 ppm (^{1}H) and 20.5 ppm (^{13}C),[5] while ^{15}N chemical shifts were referenced using glycine at −347.54 ppm. ^{1}H chemical shifts were corrected for the scaling factor due to homonuclear decoupling, which was determined using ^{1}H 1D spectra acquired under fast spinning on a Bruker 900 spectrometer. Post-processing was done using Topspin 3.5 or 3.6.1.

III. Assignment of experimental NMR spectra

The assignment of ^{13}C and ^{1}H chemical shifts for flutamide, flufenamic acid and cocaine was taken from the paper by M. Baias *et al.*[6]

The assignment of the ^{13}C spectra of ampicillin has been done by Clayden *et al.*[7] and then revised by Antzutkin *et al.*[8], but as the above authors mentioned, the assignment remains ambiguous, and so we revised it. To assign the ^{13}C NMR spectra at natural abundance a ^{13}C-^{13}C INADEQUATE experiment was done. To assign the ^{1}H directly attached to ^{13}C, the ^{1}H-^{13}C HETCOR spectra were used. To assign the ^{1}H directly attached to ^{15}N, a ^{1}H-^{15}N HETCOR experiment was done, which also helped for the assignment of ^{15}N resonances. To distinguish the ^{15}N chemical shifts belonging to NH and NH_3 resonances, a ^{15}N CP-MAT experiment was done, from which it was possible to tell that the NH_3 resonance corresponds to the peak with negligible chemical shift anisotropy due to the fast exchange of the three attached ^{1}H atoms. The assignment was cross-validated by comparing the experimental chemical shifts to shifts calculated with the GIPAW DFT method using the XRD crystal structure, albeit with optimized hydrogen positions.

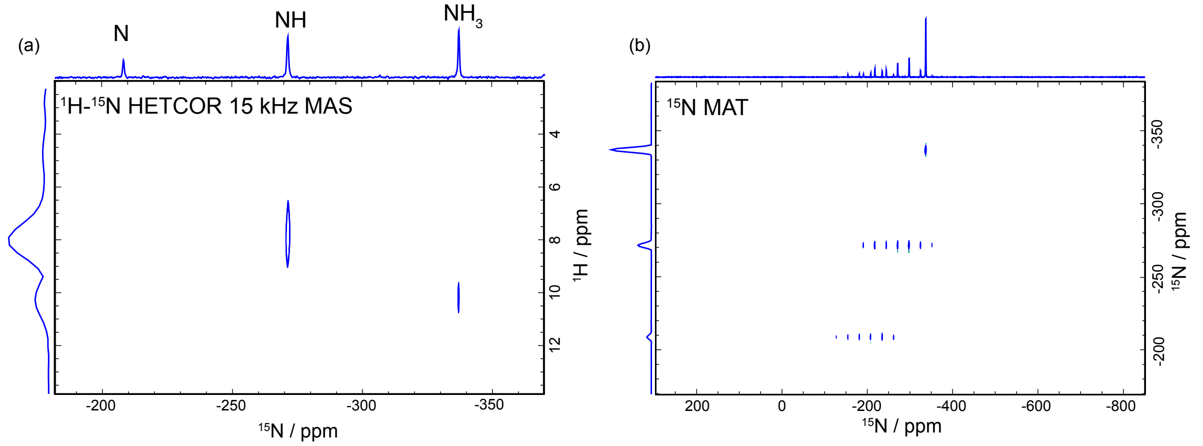

Figure S1 ^{15}N spectra of ampicillin used for the ^{1}H and ^{15}N assignments. **(a)** ^{1}H-^{15}N HETCOR spectra of ampicillin measured at 16.4 T and 15 kHz MAS. **(b)** ^{15}N MAT spectra of ampicillin at 16.4 T.

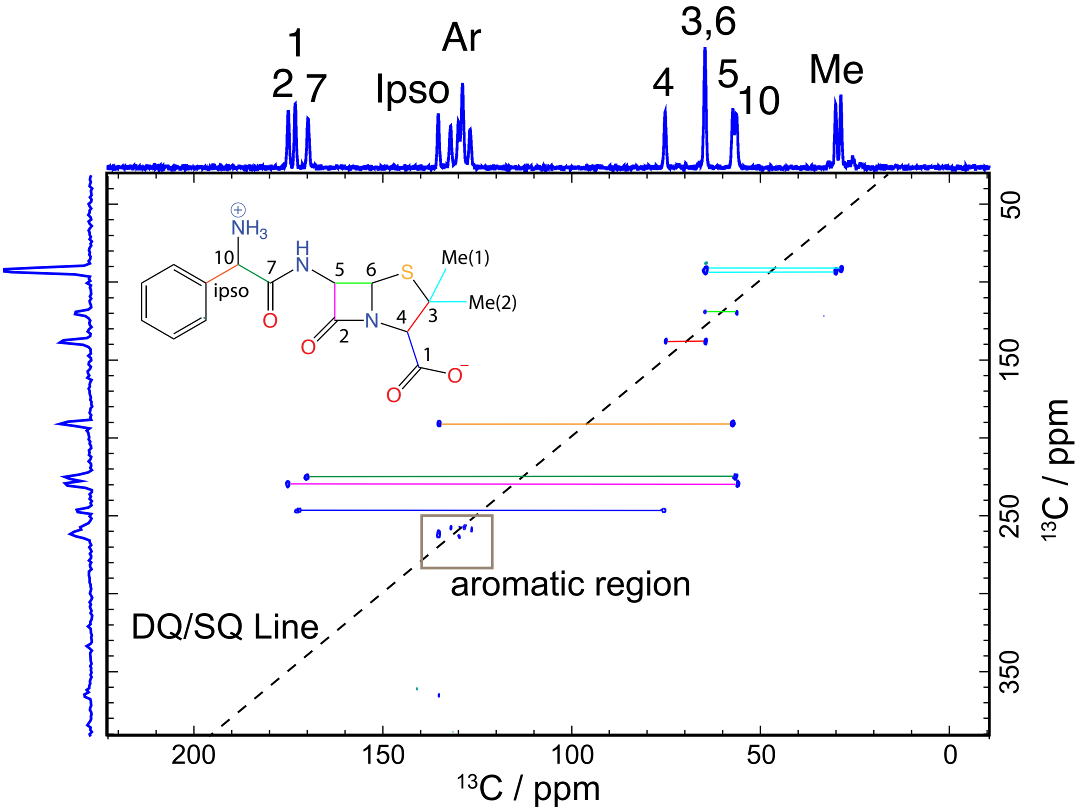

Figure S2 ^{13}C-^{13}C INADEQUATE spectra of ampicillin used for the ^{13}HC assignments, measured at 11.7 T and 13 kHz MAS.

IV. Experimental chemical shifts

Label	^{1}H, ppm	^{13}C, ppm
1	3.5	66.0
2	3.5	50.2
3	5.5	66.7
4	3.3	36.7
5	3.4	62.6
6	3.4	25.6
7	2.4	25.6
8	-	165.9
Ar	7.8	129.4
Ar (ipso)	-	134.5
15	-	172.2
16	3.5	50.2
17	1.2	41.5

Table S1 Cocaine experimental chemical shifts.

Label	^{1}H, ppm	^{13}C, ppm
1	-	149.3
2	-	109.7
3	8.3	133.0
4	6.0	117.2
5	5.4	136.3
6	6.8	112.0
7	-	175.0
8	9.6	-
9	-6.6	-
10	-	139.9
11	6.9	121.7
12	-	131.7
13	6.2	119.8
14	5.9	129.5
15	7.3	128.1
16	-	124.1

Table S2 Flufenamic acid experimental chemical shifts.

Label	^{1}H, ppm	^{13}C, ppm
1	-	145.4
2	-	124.5
3	7.9	130.9
4	-	140.9
5	9.9	124.5
6	7.1	116.7
7	-	122.0
8	8.0	-
9	-	176.1
10	2.3	35.7
11	1.3	17.7
12	1.3	21.7

Table S3 Flutamide experimental chemical shifts.

Label	^{1}H, ppm	^{13}C, ppm	^{15}N, ppm
Me_1	0.6	30.1	-
Me_2	1.6	28.9	-
4	4.0	75.3	-
10	4.8	57.4	-
6	5.2	64.8	-
Ar	5.4	128.3	-
5	6.6	56.5	-
Ar	7.1	129.0	-
Ar	7.2	132.0	-
Ar	7.3	129.9	-
Ar	7.6	126.9	-
N	-	-	Around -210
NH	7.5	-	Around -270
NH_3	10	-	Around -340
3	-	64.8	
Ar(ipso)	-	135.4	
7	-	169.8	
1	-	173.2	
2	-	175.0	

Table S4 Ampicillin experimental chemical shifts.

V. Signal to Noise analysis

The signal to noise ratio (SNR) extraction and analysis were done using the Signals extracted directly from TopSpin 4.0.5 in text file format together with a home-written python script. The SNR was extracted as:

$$SNR = maxval(S)/(2 * noise),$$

where maxval(S) is the maximum intensity at a given ^{1}H and ^{13}C chemical shifts coordinate ±0.2 ppm.- Note, that after a first extraction of maxval(S) the ^{1}H and ^{13}C coordinates were centered above maxval(S) and a refined maxval(S) was extracted.
The noise was extracted as the variance of the intensity for 100 areas ($0.4 \times 0.4\ ppm$) within the spectra. The initial 10 noise-areas were chosen manually, as to not contain any cross-peaks. The subsequent 90 noise-areas were chosen at random and were included in the noise intensity if the maximum signal intensity within the random area was less-than or-equal to two times the maximum signal intensity in an area previously selected. **Figure S3a** shows the extracted SNR of all ^{1}H-^{13}C HETCOR spectra for cocaine, flufenamic acid and flutamide against the corresponding inter-atomic distance.
First, we normalise each cross-peak by the number of protons. For this we estimate the number of protons for a given cross-peak in a spectrum by the maximum signal intensity at the given frequency, which is given from the maximum SNR at a given ^{1}H coordinate. In a next step, we take into account the difference in sensitivity between the spectra, due to the specific experimental setups, by normalising each cross-peak with respect to the maximal proton-normalised SNR per spectrum. This leads to a normalised SNR per ^{1}H, which is comparable across all experiments and is shown in **Figure S3b.**

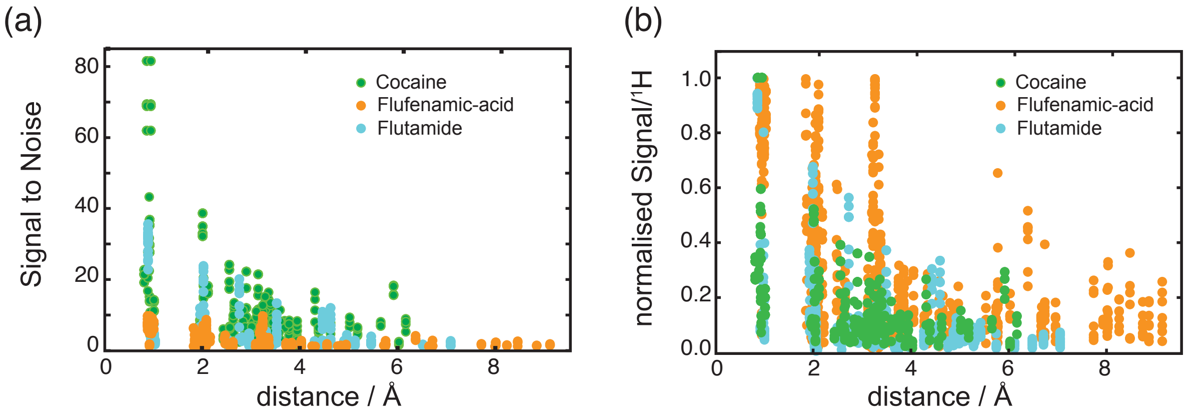

Figure S3 Signal intensity of ^{1}H-^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for cocaine (green), flufenamic acid (orange) and flutamide (cyan). **(a)** The SNR is extracted directly for all ^{1}H-^{13}C HETCOR at different contact-times and different experimental setups. **(b)** The normalised SNR per ^{1}H allows a direct comparison across different experimental setups and for cross-peaks corresponding to a different number of protons.

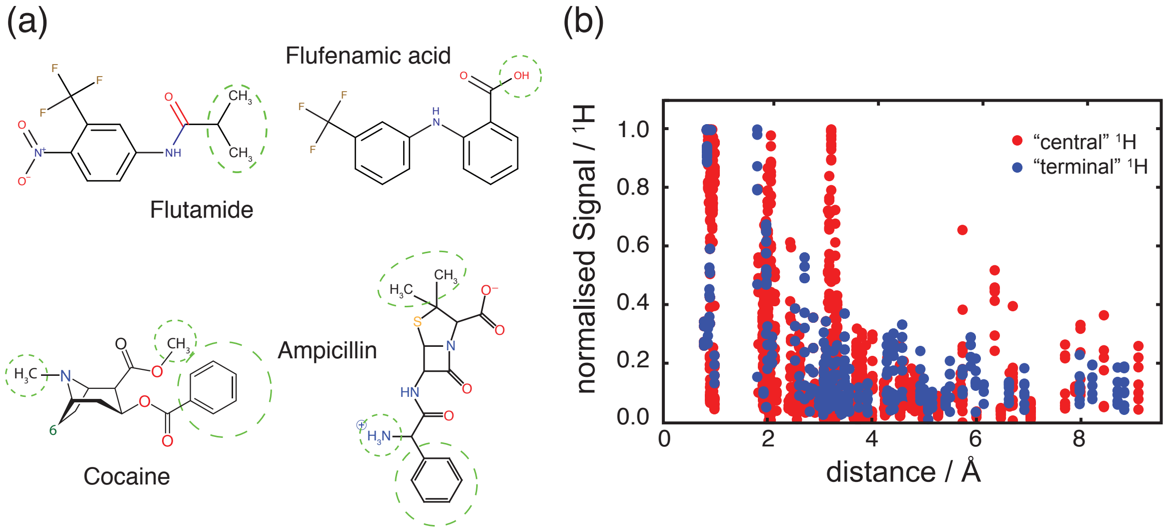

Figure S4 (a) Illustration of 'terminal' protons, which contribute to eventual conformational constraints. **(b)** normalised SNR of ^{1}H-^{13}C HETCOR cross-peaks plotted against the corresponding interatomic distance for 'centre' protons (red) and 'terminal' protons (blue), which are used to generate conformational constraints.

Molecule	‘terminal’ ^{1}H
Cocaine	Ar
	16
	17
Flufenamic acid	9
Flutamide	10
	11
	12
Ampicillin	Ar
	NH_3
	Me(1)
	Me(2)

Table S5 Protons contributing to conformational constraints for cocaine, flufenamic acid, flutamide and ampicillin

VI. Gas-phase conformer generation

For cocaine, flutamide and flufenamic acid, the CSP conformers and crystal structures were generated as described by M. Baias *et al.*[6] The coordinates of all conformers are given as separate ‘xyz’-files.

For ampicillin, we generated as complete and unbiased a set of gas phase conformers as possible using a low-mode conformational search (LMCS) method,[9-10] as implemented in MacroModel.[11] Energies were calculated during the conformer search using the OPLS3 force field.[12] The only prior knowledge used was that the molecule was present in the zwitterionic configuration throughout the conformer search. Minimum and maximum move distances of 3 and 6 Å were applied and 12,000 search steps were performed (2,000 per flexible dihedral angle). Duplicate molecular geometries were identified and removed using an all-atom RMS deviation of atomic positions, with a 0.05 Å tolerance.

All conformers were re-optimized in Gaussian09 using dispersion-corrected density functional theory (DFT-D) at the B3LYP/6-311G** level of theory with the D3BJ dispersion correction.[13] The N-H bond lengths at the amino nitrogen atom were constrained to 1.035 Å to keep the molecule in its zwitterionic form. Without this constraint, a hydrogen atom transfers from the amino to the carboxyl group during DFT reoptimization of many of the conformers. Importantly, the resulting non-zwitterionic conformers are not relevant to the known polymorphs of ampicillin.

In analysing the conformers resulting from the search, we found that the configuration around chiral centres could be reversed during the LMCS search. Therefore, all possible diastereomers of ampicillin were found to be present in the results. All conformers of a different diastereomer to that of interest were removed from the conformational ensemble before CSP was undertaken.

The coordinates of all the conformations are given as separate ‘xyz’-files.

VII. Sketch-map analysis

The cluster generation and analysis was performed with home-written Python and MATLAB codes and using the sketch-map package.[14-17] The sketch-map parameters are given **Table S5**. They were chosen following the procedure described in Ceriotti et al.[15] and the tutorial on sketchmap.org. The sketch-map analysis was not sensitive to small variations in the chosen parameters, as was already noted in the references.[15-17] As starting point for the sketch-map analysis we used all dihedral angles, not containing protons, over the full 2π range. This gives 55, 47, 31 and 35 dihedral angles for ampicillin, cocaine, flutamide and flufenamic acid, within a range of $-\pi$ to π.

Structure	$\Sigma = \sigma$	A	*B*	*a*	b
Cocaine	13	4	4	1	2
Ampicillin	6	2	2	1	1
Flutamide	6	3	3	1	1
Flufenamic Acid	6	2	2	1	1

Table S6 Sketch-map parameters for all compounds.

Ampicillin

The gas-phase CSP conformer ensemble of ampicillin contains 16 locally stable conformations (after DFT-D geometry optimization). The conformers are labeled according to increasing force-field energy. Conformer **14** is the most similar to the conformer in the crystal and resulted in the correct crystal structure after the remaining CSP procedure. **Figure S5** shows the sketch-map analysis of the ampicillin gas-phase ensemble.

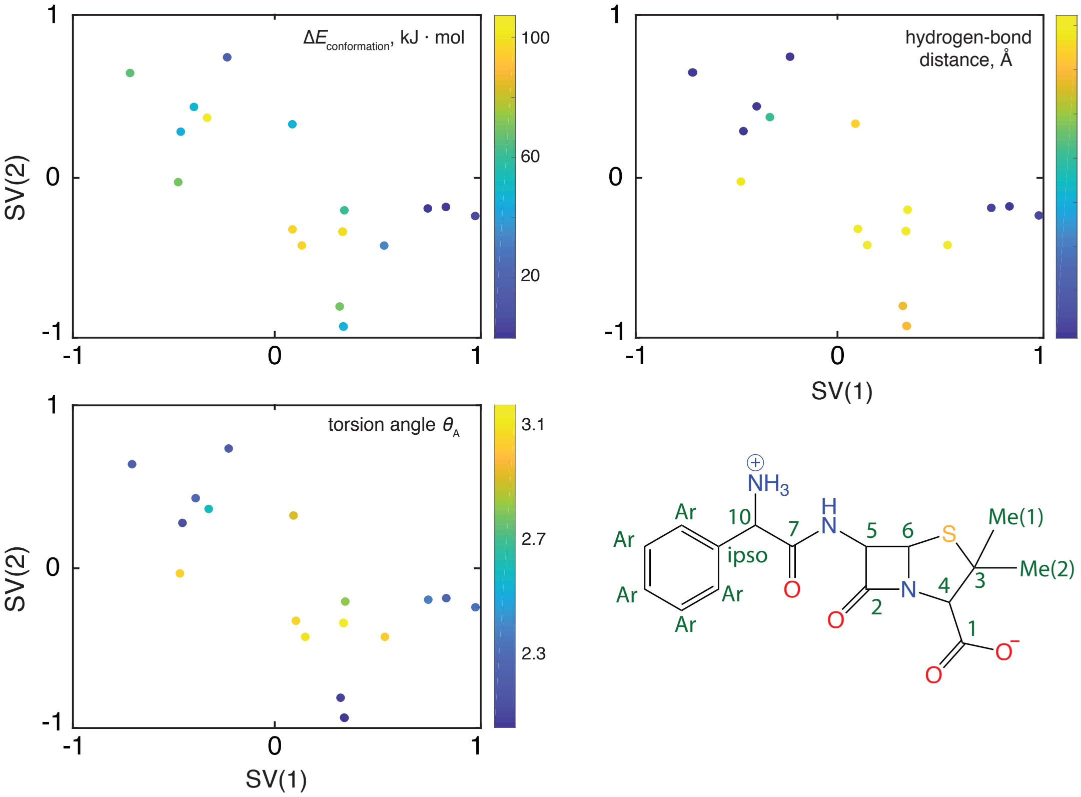

Figure S5 Sketch-map representation of the locally stable ampicillin conformers and the conformer from the single crystal XRD structure. To show the extent of the sub-clustering the panels are coloured according to different molecular properties. **Top left** shows the difference in conformational energy ($\Delta E_{\text{conformation}}$). **Top right** shows the shortest intra-molecular hydrogen-bond distance between either NH_3 or NH and the carboxyl group. **Bottom left** shows the torsion angle θ_A, which is defined as the torsion angle between C_{10}-C_7-N(H)-(N)H. In general, the clustering seems to correspond to conformational changes along the C_{ipso}-C_{10}-C_7-N(H)-C_5 chain and to relative changes between the methyl and carboxyl groups. **Bottom right,** shows the 2D structure of ampicillin with the labelling scheme used

Cocaine

The gas-phase CSP conformer ensemble of cocaine contains 27 locally stable conformations (after DFT-D geometry optimization). The conformers are labeled according to increasing force-field energy. Conformer **2** resulted in the correct crystal structure after the remaining CSP procedure.[6] **Figure S6** shows the sketch-map representation of the locally stable cocaine conformers. The main changes along the sketch-map principle components are rotations of the ester group (along SV(1)) and rotations within the methylamine group (along SV(2)).

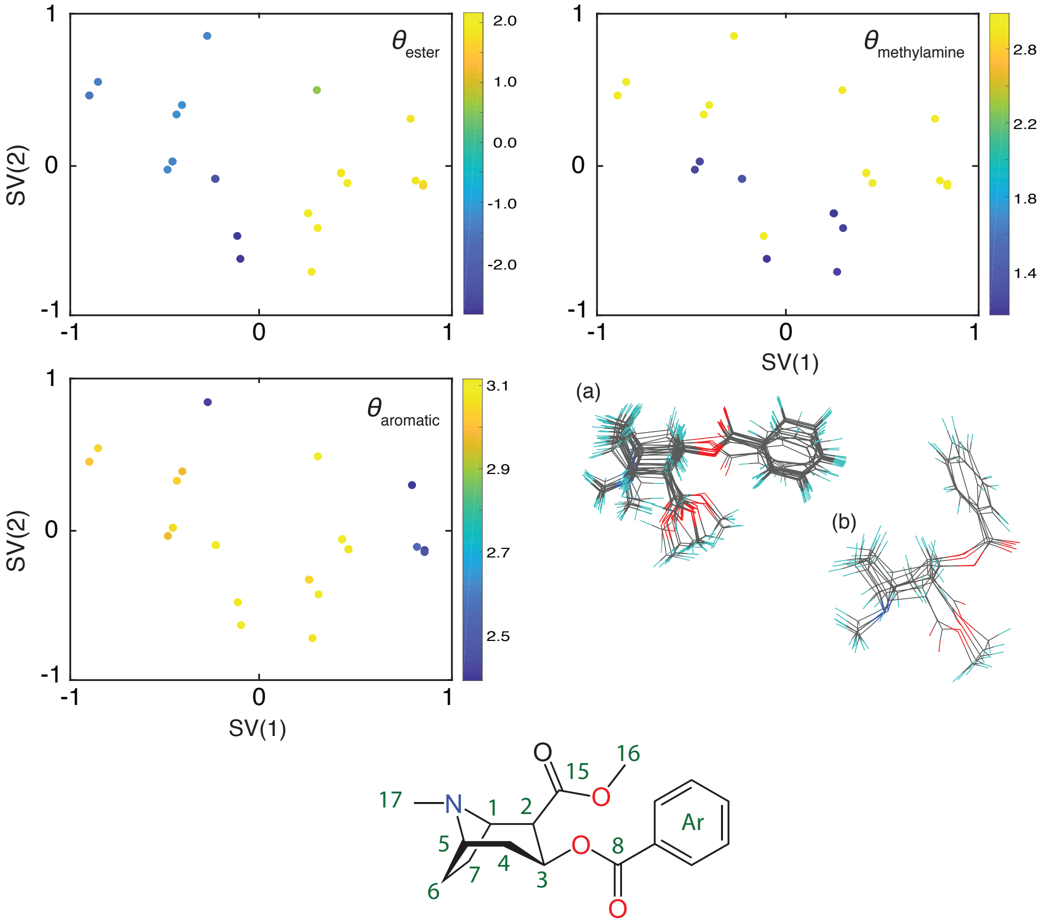

Figure S6 Top) Sketch-map representation of the locally stable cocaine conformations. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{ester} is defined as the C1-C2-C15-O4 torsion angle and reports on rotations of the ester group. $\theta_{methylamine}$ is defined as the C2-C1-N-C17 torsion angle and reports on rotations of the methyl group attached to the nitrogen. $\theta_{aromatic}$ is defined as the C(ortho)-C(ipso)-C8-O2 torsion angle and specifies the orientation of the phenyl ring plane relative to the nearby carboxylate group plane. The lower right panel shows the overlapped conformation with the aromatic ring roughly in the plane defined by the carboxylate group **(a)** and roughly perpendicular to the carboxylate group plane **(b)**. **Bottom)** 2D structure of cocaine with the labelling scheme used.

Flutamide

The gas-phase CSP conformer ensemble of flutamide contains 15 locally stable conformations (after DFT-D geometry optimization). Of those, 7 are in the trans and 8 in the cis conformation with respect to the amide group. The conformers are labeled according to increasing force-field energy. Conformer **1** resulted in the correct crystal structure after the remaining CSP procedure.[6] **Figure S7** shows the sketch-map representation of the locally stable flutamide conformers. The sketch-map representation shows a relatively distinct clustering along the sketch-map axes, which correspond to the cis and trans conformations and rotations of the methyl groups. The SV(2) axis also partially corresponds to rotations of the aromatic ring.

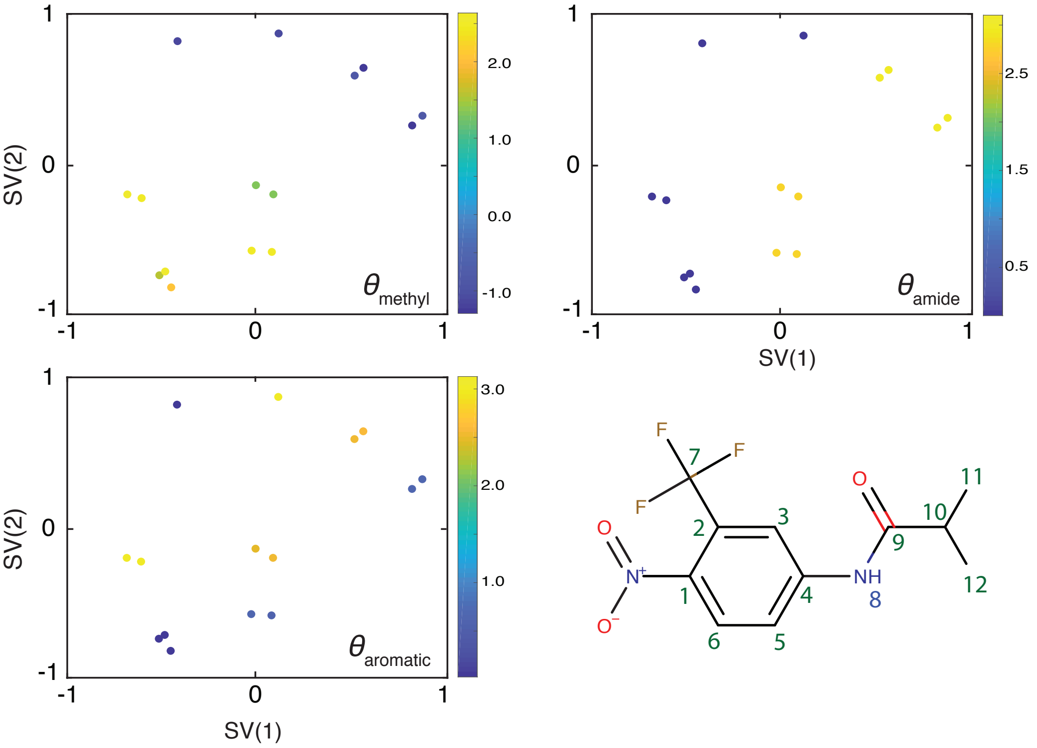

Figure S7 Sketch-map representation of the gas-phase flutamide conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{methyl} is defined as the C11-C10-C9-N(H) torsion angle and reports on rotations of the methyl carbons. θ_{amide} is defined as the C4-N(H)-C9-O1 torsion angle and reports on the amide conformation. θ_{aromatic} is defined as the C3-C4-N(H)-C9 torsion angle and reports on rotations of the aromatic group. The lower right panel shows the 2D structure of flutamide with the labelling scheme used.

Flufenamic acid

The initial CSP conformer ensemble of flufenamic acid contains 26 locally stable conformations (after DFT-D geometry optimization). The conformer **3** resulted in the correct crystal structure after the remaining CSP procedure.[6] **Figure S8** shows the sketch-map representation of the flutamide gas-phase conformer ensemble. The main changes along the sketch-map principle components correspond to rotations of the carboxyl group (along SV(1)) and rotations of the two aromatic groups (along SV(2)).

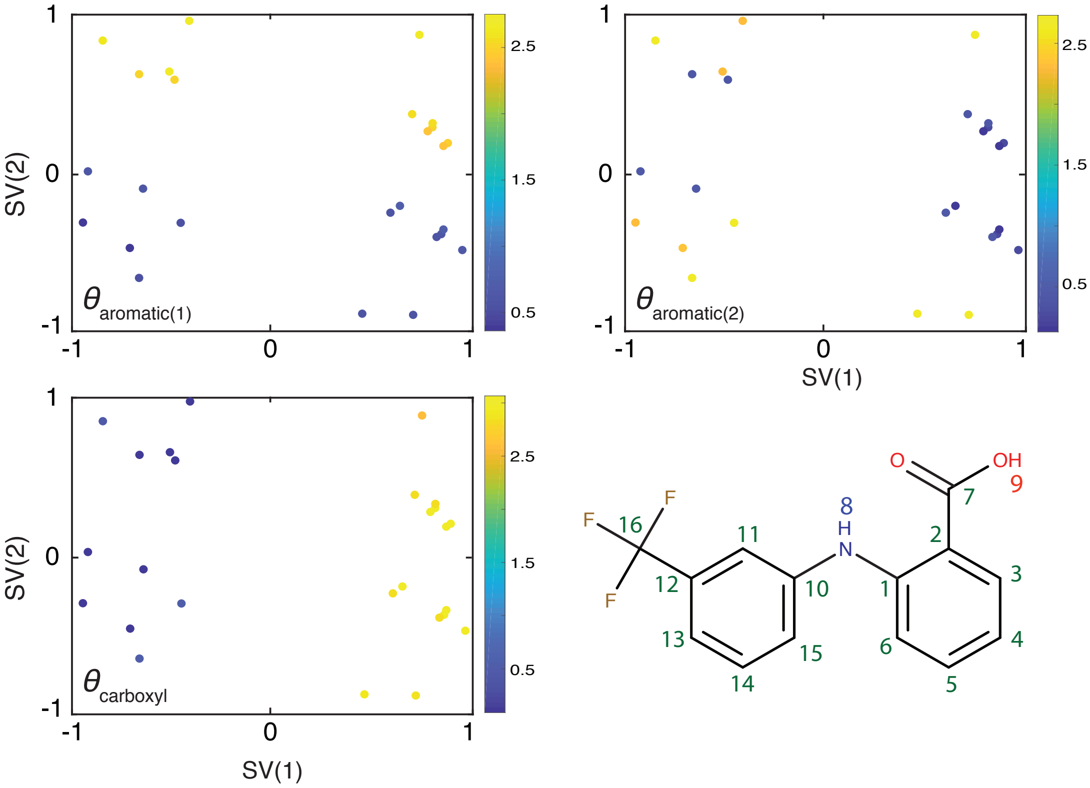

Figure S8 Sketch-map representation of the gas-phase flufenamic acid conformers. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. $\theta_{\mathrm{aromatic(1)}}$ is defined as the C15-C10-N(H)-H(N) torsion angle and reports on rotations of aromatic ring with the attached trifluormethyl. $\theta_{\mathrm{aromatic(2)}}$ is defined as the C7-C2-N(H)-H(N) torsion angle and reports on rotations of aromatic ring with the attached carboxyl. $\theta_{\mathrm{carboxyl}}$ is defined as the C1-C2-C7-O(H) torsion angle and reports on rotations of the carboxyl group. The lower right panel shows the 2D structure of flufenamic acid with the labelling scheme used.

VIII. Parametrization of the constraints

It is not a priori clear as to what the threshold distance *X* should but in general we expect ^{1}H-^{13}C HETCOR cross-peaks in solid-state NMR spectra to appear for all interatomic distances of up to 3.5 Å. Here, we investigate the use of threshold distances (*X*) from 2.0 to 5.0 Å in steps of 0.5 Å and for S_{norm} cut-off values from 0.08 to 0.22 in steps of 0.02 for the polymorphs of cocaine, flutamide, flufenamic acid. **Figure S9** shows the set of successful parameters for each molecule individually.

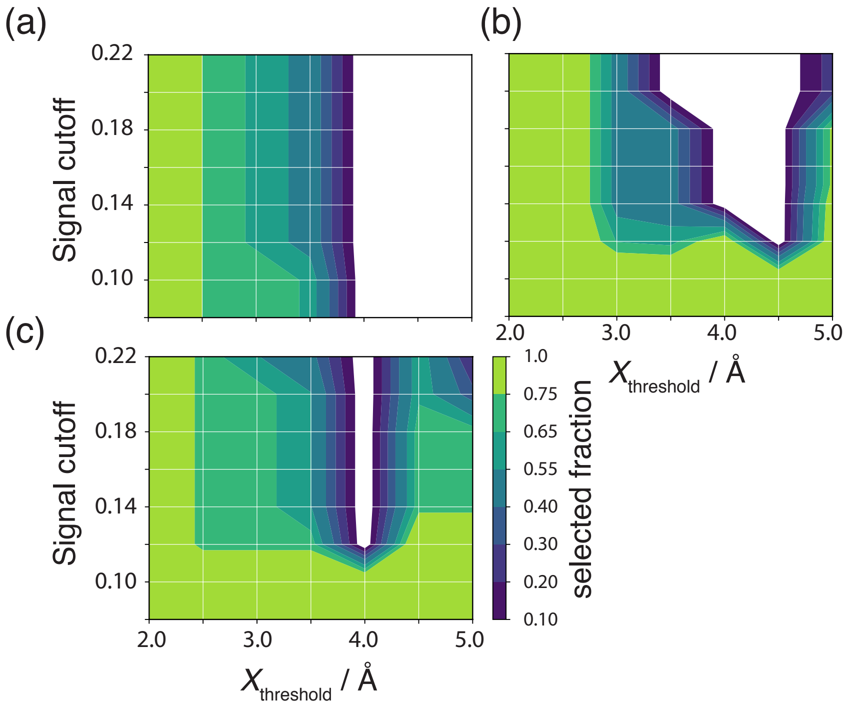

Figure S9 Grid search results of the threshold distance (*X*) and normalized SNR cut-off values (S_{norm}) for **(a)** flutamide, **(b)** cocaine and **(c)** flufenamic acid. The colour-map shows the percentage of selected structures from within the conformer ensemble. The white area indicates the region where the correct conformer was not selected. Optimal selection parameters should select the smallest conformer sub-ensemble, while still containing the correct structure. This corresponds to the dark blue regions within the different panels.

IX. Conformer selection

The ensemble selection was done with home-written Python codes. The HETCOR cross peaks below a S_{norm} value of 0.14 were interpreted as hydrogen-carbon distances greater than 3.5 Å (the "threshold" distance, X). For each conformation, the number of fulfilled constraints was counted and the conformers were sorted in decreasing order.

Flutamide.

Conformer selection for flutamide was done based on constraints from multiple HETCOR NMR experiments, with variable contact times of 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.0 ms. The ^{1}H and ^{13}C cross-peaks from the two methyl groups could not be distinguished. Also, the ^{1}H cross peaks from H3 and H8 as well as the ^{13}C cross peaks from C5 and C2 are too close and thus indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. The HETCOR cross-peaks are listed in the file "flutamide_SN_auto.csv". All the conformers, sorted by the number of experimental constraints satisfied are given in the Excel file "ensemble_selection.xlsx".

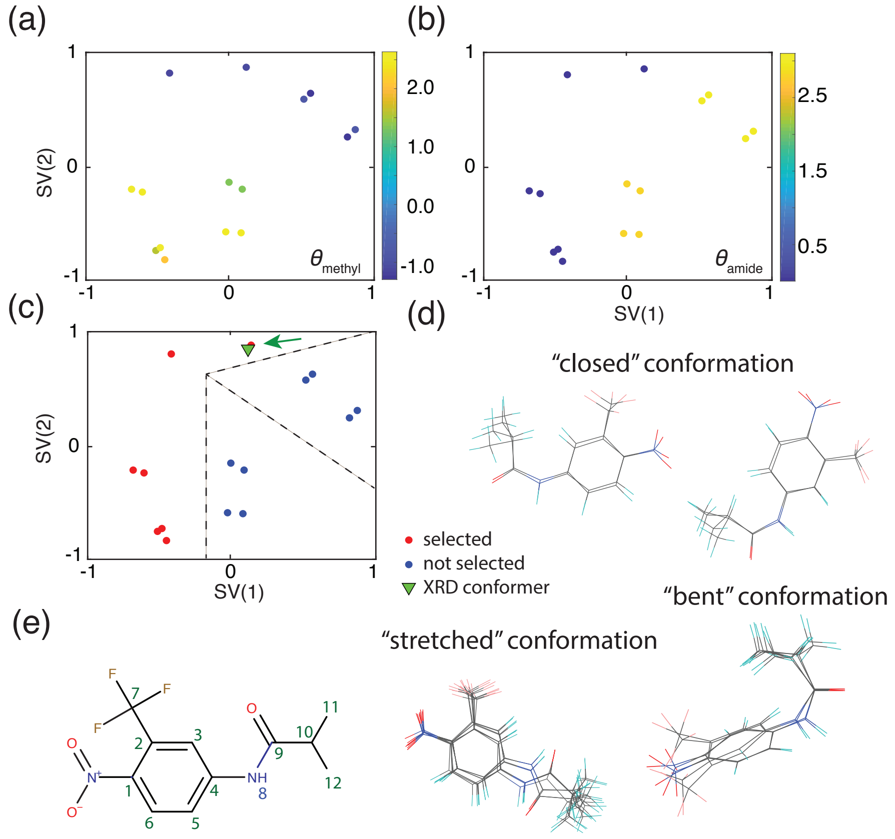

Figure S10 (a-b) Sketch-map representation of the gas-phase flutamide conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. θ_{methyl} is defined as the C11-C10-C9-N(H) torsion angle and reports on rotations of the methyl groups. θ_{amide} is defined as the C4-N(H)-C9-O1 torsion angle and reports on the amide conformation. **(c)** Sketch-map projection of the gas-phase flutamide ensemble. Red dots represent the conformers with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. The black dashed lines indicate the regions where the different conformer sub-ensembles, shown in (d) are located. **(d)** Overlay of the structures within the different sketch-map clusters. The "stretched" conformations correspond to the trans conformers and are all selected. The "bent" and "closed" conformations correspond to the cis conformers and are not selected. **(e)** 2D structure of flutamide with the labelling scheme used.

Cocaine.

The ^{1}H-^{13}C HETCOR NMR experiments were performed on cocaine at variable contact times of 0.5, 0.75, 1.0 and 1.5 ms. The ^{1}H and ^{13}C cross-peaks from the aromatic group could not be distinguished. Also, the ^{13}C cross-peaks from C6 and C7, the ^{13}C cross-peaks from C2 and C16, as well as the ^{1}H cross-peaks from H1, H2, H4, H5, and H6 were too close and hence indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. The HETCOR cross-peaks are listed in the file "cocaine_SN_auto.csv". All the conformers sorted by number of experimental constraints satisfied are given in the Excel file "ensemble_selection.xlsx".

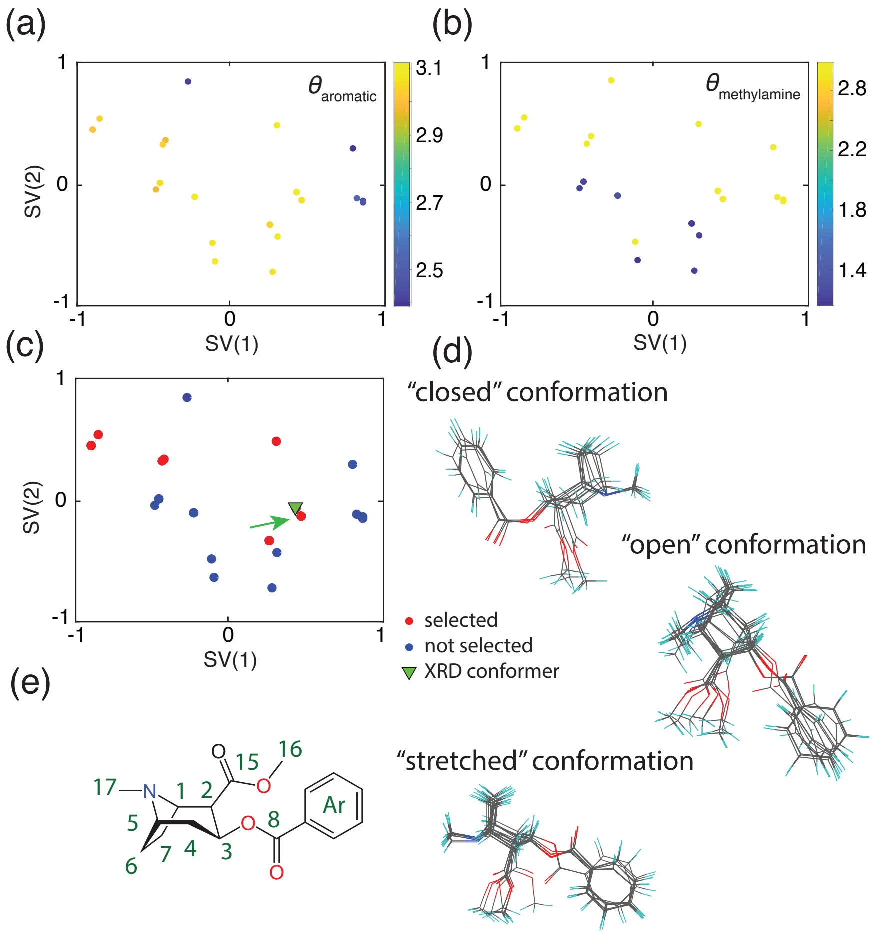

Figure S11 (a-b) Sketch-map representation of locally stable cocaine conformers. To show the extent of the sub-clustering, the panels are coloured according to different torsion angles reporting on different torsion angle values in the molecule. $\theta_{\text{methylamine}}$ is defined as the C2-C1-N-C17 torsion angle and reports on rotations of the methyl group attached to the nitrogen. θ_{aromatic} is defined as the C(ortho)-C(ipso)-C8-O2 torsion angle and reports on flips of the aromatic group. **(c)** Sketch-map projection of the gas-phase cocaine ensemble. Red dots represent the structures with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. **(d)** Overlay of the structures within the different sketch-map clusters. The "stretched" conformations correspond to the selected conformers. The "closed" conformations contain different θ_{aromatic} torsional angle and are not selected. The "open" conformation contains a different $\theta_{\text{methylamine}}$ torsional angle and are not selected. **(e)** 2D structure of cocaine with the labelling scheme used.

Flufenamic Acid.

Conformer selection for flufenamic acid was done based on constraints from multiple HETCOR NMR experiments, with variable contact times of 0.1, 0.5, 1.0, 1.5, 3.0 and 3.5 ms. The ^{1}H cross-peaks from H4, H13 and H14 as well as the ^{1}H cross-peaks from H6, H11 and H15 are too close and thus indistinguishable. Therefore, if a cross-peak was seen it was attributed to all atoms in the given group. The HETCOR cross-peaks are listed in the file "flufenamic_acid_SN_auto.csv". All the conformers, sorted by the number of experimental constraints satisfied, are given in the Excel file "ensemble_selection.xlsx".

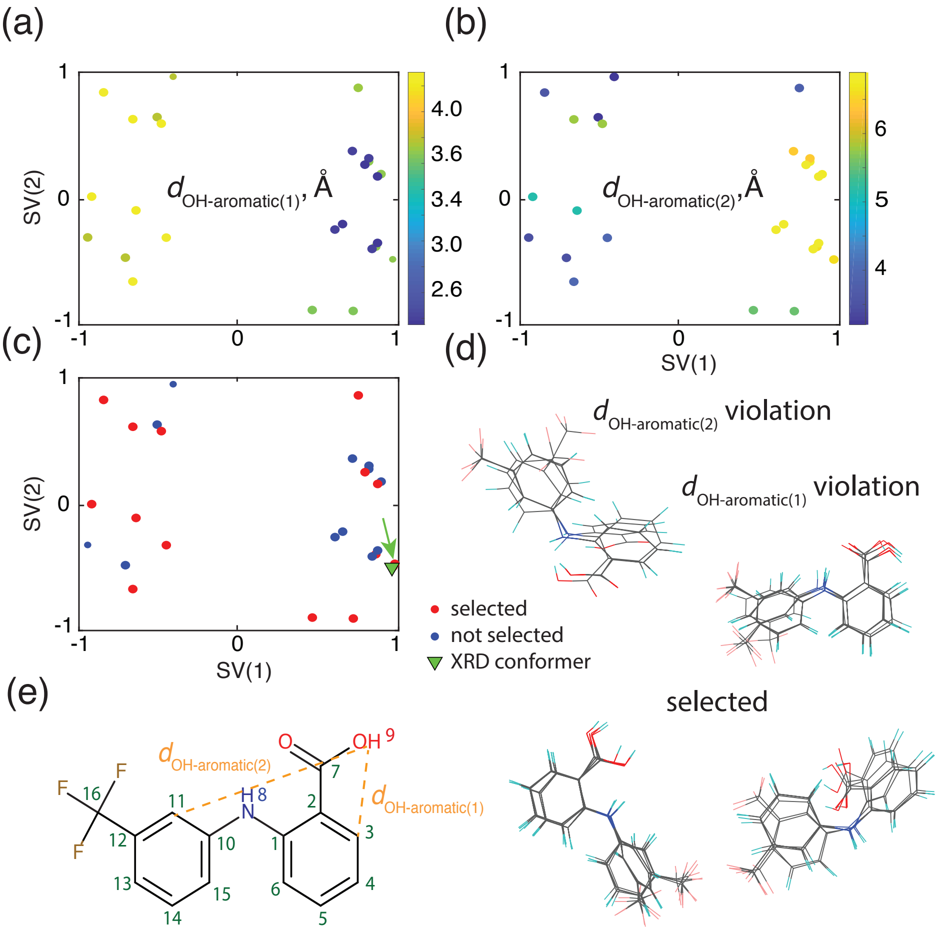

Figure S12 (a-b) Sketch-map representation of the gas-phase flufenamic acid conformer ensemble. To show the extent of the sub-clustering, the panels are coloured according to the distance [Å] between the OH group and the two aromatic rings. The distance is expressed as the distance between the carboxyl proton and C3/C11 (as shown in **e**). **(c)** Sketch-map projection of the gas-phase flufenamic acid conformer ensemble. Red dots represent the conformers with the lowest number of constraint violations, and are thus selected. The green triangle shows the conformer present in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer, which resulted in the correct crystal structure after the CSP procedure. **(d)** Overlay of the structures within the different sketch-map clusters. **(e)** 2D structure of flufenamic acid with the used labelling scheme.

Ampicillin

The conformer selection for ampicillin was done using constraints from ^{1}H-^{13}C HETCOR NMR experiments with variable contact times of 0.1, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75 and 2.25 ms. The ^{1}H and ^{13}C cross-peaks from the two methyl groups could not be distinguished. Also, the ^{1}H cross-peaks from Ar2-6, H5 and NH, the ^{1}H cross-peaks from Ar1, H10 and H6, the ^{13}C cross-peaks from C3 and C6 as well as the ^{13}C cross-peaks from Ar1-5 were too close and hence indistinguishable. Therefore, if a cross-peak was observed it was attributed to all atoms in the given group. The HETCOR cross-peaks are listed in the file "ampicillin_SN_auto.csv". All the conformers, sorted by the number of experimental constraints satisfied, are given in the Excel file "ensemble_selection.xlsx".

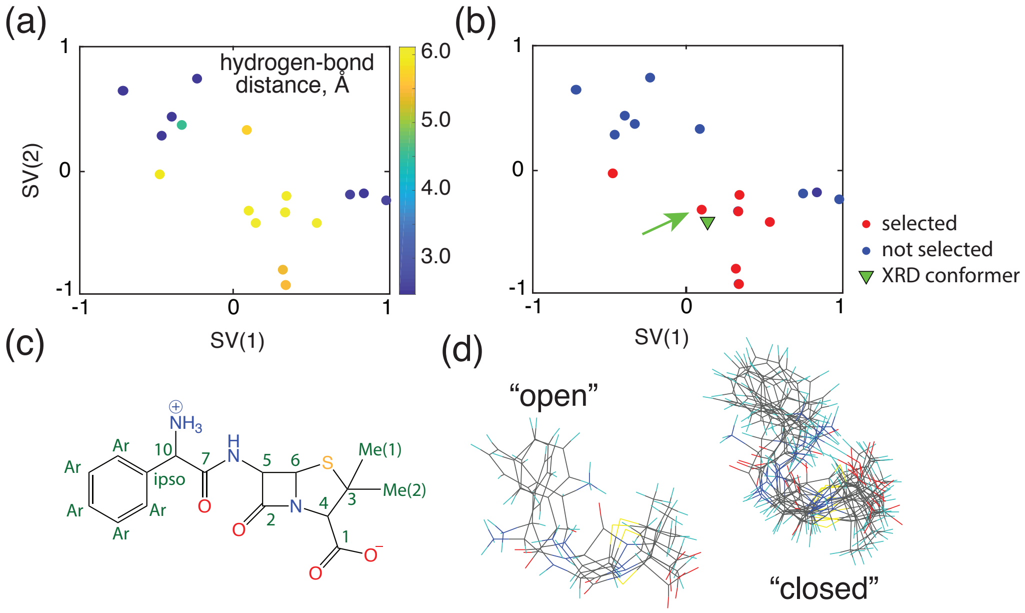

Figure S13 **(a)** Sketch-map representation of the locally stable ampicillin conformations. To show the extent of the sub-clustering, the panel is coloured according to the shortest intra-molecular hydrogen-bond distance [Å] between either NH_3 or NH and the carboxyl group. **(b)** Sketch-map projection of the gas-phase ampicillin ensemble. Red dots represent structures with the lowest number of constraint violations, and are therefore selected. The green triangle shows the conformer found in the XRD-determined crystal structure. The green arrow points to the gas-phase conformer that resulted in the correct crystal structure after the CSP procedure. **(c)** 2D structure of ampicillin with the labelling scheme used. **(d)** Overlay of the structures within the different sketch-map clusters. The "open" conformations correspond conformers without an intra-molecular hydrogen bond and were selected. The "closed" conformations mostly contain an intra-molecular hydrogen bond and were not selected.

X. Ampicillin crystal structure generation

From the 7 selected conformations (AMCILL_OPLS3_5, AMCILL_OPLS3_7, AMCILL_OPLS3_10, AMCILL_OPLS3_12, AMCILL_OPLS3_13, AMCILL_OPLS3_14 and AMCILL_OPLS3_15, provided in the Supplementary Material) a set of crystal structures was generated using a low-discrepancy, quasi-random search of crystal packing variables using the GLEE (Global Lattice Energy Explorer) code.[18] Each space group considered is sampled separately, by generating trial structures with unit cell dimensions, molecular positions and orientations sampled using a low discrepancy method. Crystal structures were generated in the 11 most commonly observed Sohncke space groups (1, 4, 5, 18, 19, 76, 78, 92, 96, 144, 145) until 2000 valid (successfully lattice energy minimized) crystal structures were generated in each space group, for each conformer.

All generated trial crystal structures were geometry-optimized using the crystal structure modelling code DMACRYS[19] with the molecular geometry fixed at the gas-phase geometry. Intermolecular interactions were evaluated using the s using atomic multipoles, up to hexadecapole on each atom, derived using a distributed multipole analysis[20] of the B3LYP/6-311G** charge density.

All predicted crystal structures within 20 kJ mol^{-1} in total (intermolecular + conformational) energy of the lowest energy structure were then re-optimized using dispersion-corrected DFT. To ensure that all selected conformers were represented in the final crystal structures, a minimum of 5 crystal structures were taken from each conformer (whether or not they fell within the lowest 20 kJ mol^{-1}). This selection resulted in a total of 75 crystal structures. These were relaxed with DFT using the Castep[21] suite, using the PBE functional, the D2 dispersion correction, a 500 eV basis set cutoff and *k*-points sampled on a Monkhorst-Pack grid to provide a maximum reciprocal point spacing of 0.04 Å^{-1}. Each crystal structure was optimized in two stages: first, with the unit cell fixed from the force field predicted crystal structure, then fully relaxed, including the unit cell and all atomic positions. The resulting structures were used as starting points for the chemical shift modelling (see below). All the relaxed structures are given as Castep[21] output files in the zip-folder "castep_output.zip".

XI. Ampicillin chemical shift calculations and crystal structure selection

Structure modelling.
Prior to the chemical shift calculations, all the trial structures and the single-crystal XRD structure of ampicillin[22] were fully relaxed, including the unit cell and all atomic positions, using the same DFT parametrization as outlined for the chemical shift calculations below.

DFT chemical shift calculation.
For the ampicillin crystal structure selection, the magnetic shielding at the ^{1}H and ^{13}C nuclei in the 75 trial crystal structures were calculated with plane-wave DFT using the GIPAW formalism[23] and the Quantum ESPRESSO suite.[24] For the GIPAW DFT calculations, the generalized-gradient-approximation (GGA) density functional PBE[25] was used. We used the ultrasoft pseudopotentials with GIPAW[26-27] reconstruction, C.pbe-n-kjpaw_psl.1.0.0.UPF, N.pbe-n-kjpaw_psl.1.0.0.UPF, H.pbe-kjpaw_psl.1.0.0.UPF, O.pbe-nl-kjpaw_psl.1.0.0.UPF and S.pbe-nl-kjpaw_psl.1.0.0.UPF from the pslibrary database [https://dalcorso.github.io/pslibrary/]http://theossrv1.epfl.ch/Main/Pseudopotentials. A wave-function energy cut-off of 100 Ry, a charge density energy cut-off of 400 Ry and a Monkhorst-Pack grid of *k*-points[28] corresponding to a maximum spacing of 0.04 Å^{-1} in the reciprocal space was used. The electron density self-consistency convergence threshold was set to 10^{-12} Ry. All Quantum ESPRESSO input and output files are given in the zip-folder "qe_output.zip".

ShiftML chemical shift calculation.
For the ampicillin crystal structure selection, the magnetic shielding of the 75 trial crystal structures were calculated using the ShiftML version described in below.

Shielding to shift conversion and RMSE calculation.
The calculated magnetic shielding was referenced to the experimental chemical shifts using the linear relationship $\sigma_{exp} = a - b\delta_{DFT}$, where the slope was fixed (i.e., $b = 1$) and the offset (a) was fit for each trial structure individually. For the ^{1}H chemical shift RMSE calculation, the methyl protons of each methyl group and the NH$_3$ protons were averaged. As it was not possible to distinguish the aromatic protons experimentally, as well as to distinguish the 2 methyl groups experimentally, the chemical shifts within each group were sorted, both for experimental and DFT chemical shifts, and then compared to each other. This was done for each crystal structure individually. The RMSE was calculated as,

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\sigma_{i,exp} - \sigma_{i,calc})^2}{N}},$$

where σ_{exp} denotes the experimental chemical shift, σ_{calc} denotes the calculated chemical shift and the index *i* runs over all protons (*N*) within the asymmetric unit.

XII. Machine-learning model (ShiftML)

The machine-learning model used to predict the ^{1}H chemical shifts follows the basic concepts behind ShiftML,[29] which are detailed in Paruzzo et al.[29] However, the original implementation of ShiftML is only able to predict ^{1}H chemical shifts of structures containing H,C,N and O atoms. Thus, we extended the training set in the following manner:

a) Starting from the CSD-61k set and including the CSD-2k training set, described in Paruzzo et. al.,[29] we used a farthest point sampling algorithm (FPS) to include an additional 1,000 training structures.

b) From the Cambridge Structural Database (CSD),[30] we extracted a set of around 22'000 molecular crystal structures containing less than 200 atoms in the unit-cell and containing H,C and S atoms as well as optionally N and O atoms (CSD-S22k). This set was curated analogously to the CSD-61k set and using a FPS algorithm we selected 546 structures from this set.

c) These three structure sets were combined to form the CSD-3k+S546 set.

d) As structures often contain redundant environments, for example due to crystal symmetries, the training set was reduced by FPS ordering the individual environments and retaining only the 65,000 most structurally diverse.

Additionally, the ShiftML model was changed to contain radially scaled smooth overlap of atomic positions (SOAP) kernels[16, 31-32] as opposed to the seven multi-scale SOAP kernels described in Paruzzo et al.[29] This change was implemented to increase the computational efficiency of the model. The parameters of the used ShiftML implementation are given in **Table S7**, using the same notation as in Paruzzo et al.[29] and Willatt et al. [32]

In order to estimate the prediction accuracy of the updated ShiftML model, we combined the CSD-500 test set from Paruzzo et al.[29] with 104 random structures extracted from the CSD-S22k set. For this combined CSD-500+S104 test set, we find a RMSE of 0.44 ppm between ^{1}H chemical shifts calculated with DFT and ShitML. This is directly comparable to the ^{1}H RMSE of 0.49 ppm reported in Paruzzo et al.[29] We ascribe the slightly lower ^{1}H chemical shift RMSE to the fact that a larger training set was used.

Note, that all the DFT calculations and all of the treatment of the training set, e.g. the detection of unusual environments, was done as described in Paruzzo et al.[29]

Atom	r_c (cutoff)	*c* (cutoff rate)	*m* (cutoff dexp)	r_0 (cutoff scale)	u_0 (central weight)	*gw* (atom sigma)	n_{max}	l_{max}	**cutoff transition width**
^{1}H	5	1	4	2.5	1.0	0.3	9	9	0.5

Table S7 Parameters used for the implemented ShiftML version.

The RMSE between the ^{1}H chemical shifts calculated with DFT and ShiftML is calculated as 0.464 pp over all the ampicillin trial structures. This agrees with the predicted ^{1}H chemical shift RMSE, calculated as 0.44 ppm **Figure S14** shows the correlation between ^{1}H magnetic shieldings calculated with DFT and ShiftML.

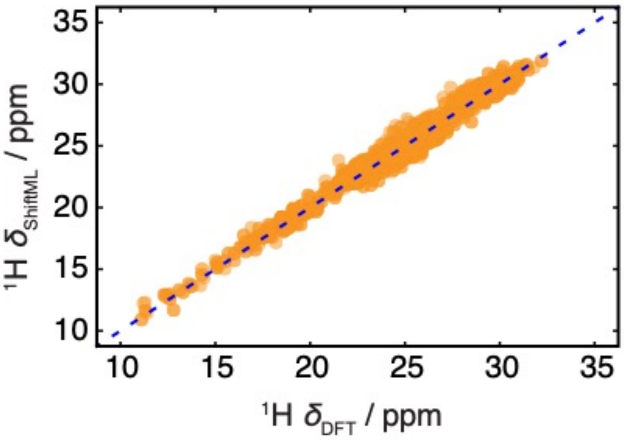

Figure S14 Scatterplot showing the correlation between ^{1}H magnetic shieldings calculated with DFT and ShiftML, with a RMSE of 0.464 ppm. The blue dotted line indicates a perfect correlation.

Figure S15 shows the RMSE between ShiftML calculated and measured ^{1}H chemical shifts together with the DFT calculated relative lattice energies for the candidate set. Note that the RMSE between experiment and the ShiftML predicted chemical shifts follows the same trends as the RMSE between experiment and the DFT calculated shifts (**Figure 6a**).

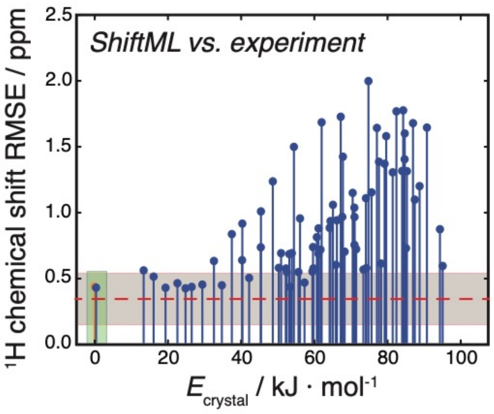

Figure S15. Comparison of crystal structure candidates. The structures are sorted according to their relative lattice energy, horizontal axis. The vertical axis shows ^1H chemical shift RMSE between ShiftML calculated and experimental chemical shifts. The orange marker shows the ^1H chemical shift RMSE for the single-crystal XRD structure. The red line shows the mean of the current error (0.346 ppm) between experimental and ShiftML calculated ^1H chemical shifts with the limits at one standard deviation (0.195 ppm) indicated as grey shaded zone, as described below.

XIII. ShiftML error estimation.

Comparison between ^1H experimental chemical shifts and ^1H chemical shifts calculated with ShiftML were carried out analysing around 150 chemical shifts obtained from 11 crystal structures. The names, IUPAC IDs, CSD reference codes (when available) and references to the experimental NMR data of the analysed crystal structures are the following:

(i) Naproxen, (2S)-2-(6-Methoxy-2-naphthyl)propanoic acid, COYRUD11, Ref.[33]
(ii) Uracil, Pyrimidine-2,4(1H,3H)-dione, URACIL, Ref. [34]
(iii) Co-crystal of 3,5-dimethylimidazole and 4,5-dimethylimidazole, Ref. [35]
(iv) Theophylline, 1,3-Dimethyl-3,7-dihydro-1H-purine-2,6-dione, BAPLOT01, Ref. [6]
(v) Anthranilic acid, AMBACO05, Refs.[36-37]
(vi) Cimetidine, CIMETD, Refs.[37-38]
(vii) Phenobarbital, PHBARB06, Refs.[37, 39]
(viii) Thymol, IPMEPL, Ref.[40]
(ix) Terbutaline hemisulfate, ZIVKAQ, Refs.[37, 41]
(x) Cocaine, methyl (1R,2R,3S,5S)-3- (benzoyloxy)-8-methyl-8-azabicyclo[3.2.1] octane-2-carboxylate, COCAIN10, Ref. [6]
(xi) AZD8329, 4-[4-(2-adamantylcarbamoyl)-5-tert-butylpyrazol-1-yl]benzoic acid, Ref.[42]

The crystal structures (i-ix) were obtained from Ref. [37], where the experimentally determined crystal structures were subjected to all-atom geometry optimization with fixed lattice parameters, as described in the reference. Crystal structures (x) and (xi) were obtained from Refs. [6] and [42] respectively. Additionally, all the used crystal structures are given in the ESI. We only used the ^1H chemical shifts from the references, which were clearly distinguishable and did not have a broad peak spanning several ppm. We used assigned chemical shift values and we account for rotational dynamics of the methyl groups by averaging the chemical shift values of the three ^1H positions to a single value for each methyl group. For chemical shifts which could not be assigned unambiguously, such as e.g. shifts from CH_2 protons, we assigned the chemical shifts on a best match basis. The calculated magnetic shieldings σ are converted to the corresponding chemical shifts δ through the relationship $\sigma_{exp} = a - b\delta_{DFT}$, where the slope (b) and the offset (a) were fit for each reference structure individually. The chemical structures, the RMSE between experimental and ShiftML predicted ^1H chemical shifts, together with the assigned experimental chemical shifts and the parameters for conversion between shieldings and shifts are shown in **Figure S16** and **Table S8**. For the entire reference set we calculate an average RMSE of 0.346 ppm and a standard deviation of 0.195 ppm.

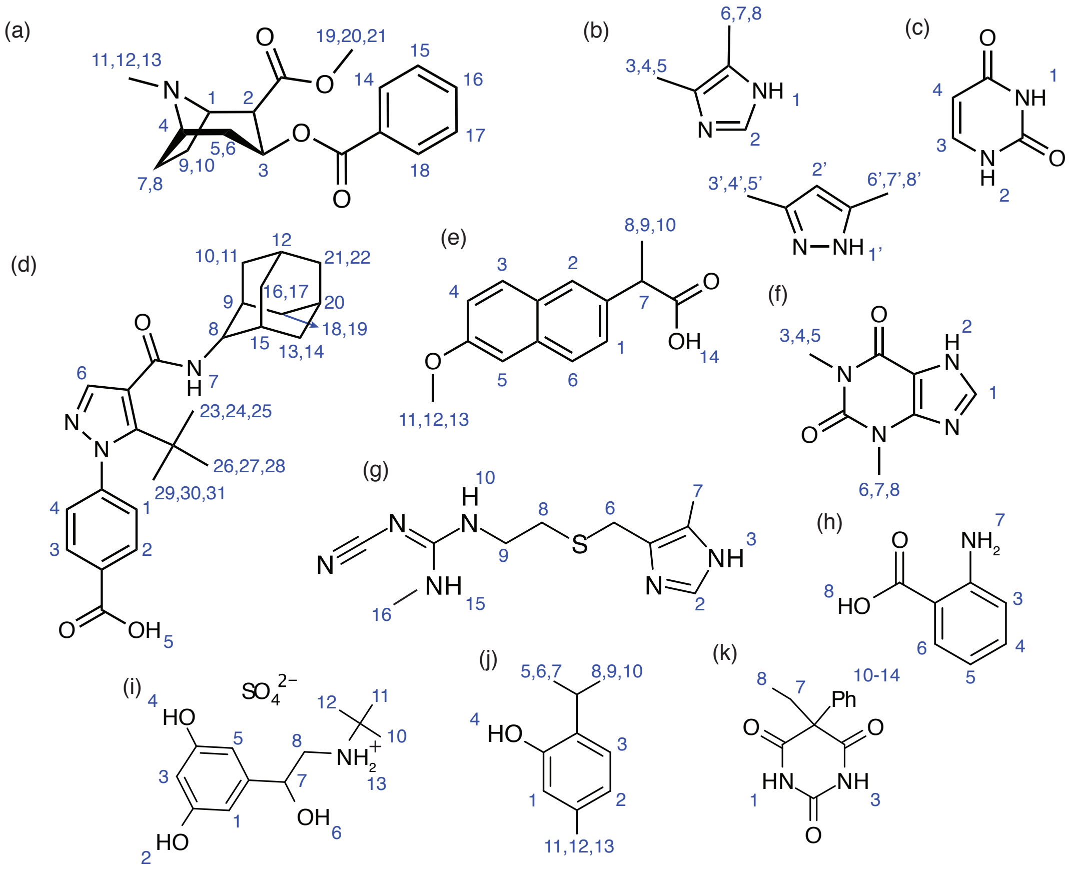

Figure S16. Chemical structures of the compounds used for experimental comparison. In order, cocaine (a), 3,5-dimethylimidazole and 4,5-dimethylimidazole (b), uracil (c), AZD8329 (d), naproxen (e), theophylline (f), cimetidine (g), anthranilic acid (h), terbutaline hemisulfate (i), thymol (j) and phenobarbital (k) and the labelling scheme used here.

Naproxen			**Uracil**		
Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)	Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
1	7	6.44	3	7.5	7.43
2	6.1	5.60	2	10.8	10.79
3	3.8	3.86	1	11.2	11.22
4	4.5	4.65	4	6	6.05
5	4.1	4.65			
6	5.9	5.48			
7	3.2	2.88			
8,9,10	1.8	1.56			
11,12,13	2.3	2.80			
14	11.5	11.75			
***a* = 4.79 ppm**	***b* = 0.81**	***RMSE* = 0.393 ppm**	***a* = 5.15 ppm**	***b* = 0.77**	***RMSE* = 0.048 ppm**

3,5-dimethylimidazole & 4,5-dimethylimidazole

Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
1’	13.0	13.25
2’	4.8	5.03
3’,4’,5’	1.4	1.12
6’,7’,8’	0.7	1.07
1	15	14.62
2	5.2	5.52
3,4,5	1.5	1.47
6,7,8	1.4	1.20
a = 4.85 ppm	***b = 0.92***	***RMSE = 0.27 ppm***

Theophylline

Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
2	14.6	14.79
1	7.7	7.10
3,4,5	3.4	3.54
6,7,8	3.4	3.40
a = 5.19 ppm	***b = 0.84***	***RMSE = 0.24 ppm***

Cocaine

Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
1	3.76	4.17
2	3.78	2.79
3	5.63	5.78
4	3.32	3.54
5	3.06	1.83
6	3.49	2.56
7	2.91	2.17
8	3.38	3.04
9	2.56	2.19
10	2.12	2.37
11,12,13	1.04	1.87
14	8.01	7.90
15	8.01	7.90
15	8.01	7.90
17	8.01	7.90
18	8.01	7.90
19,20,21	3.78	4.27
a = 5.88 ppm	***b = 1.05***	***RMSE = 0.59 ppm***

AZD8329

Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
1	6.92	6.54
2	8.69	8.30
3	9.01	8.74
4	8.47	7.64
5	15.37	14.90
6	7.73	8.04
7	9.64	10.70
8	2.90	2.72
9	1.78	2.03
10	1.88	2.28
11	1.88	2.28
12	1.8	1.99
13	1.6	1.48
14	0.44	1.21
15	1.54	1.71
16	1.88	2.10
17	1.88	2.10
18	0.8	1.39
19	0.8	1.39
20	1	1.85
21	1.74	1.75
22	1.74	1.75
23,24,25	0.73	0.39
26,27,28	0.73	0.83
29,30,31	0.73	-0.16
a = 5.40 ppm	***b = 1.06***	***RMSE = 0.50 ppm***

Cimetidine			Anthranilic acid		
Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)	Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
2	7.64	7.55	Aromatic (1)	5.8	5.74
3	11.84	11.55	Aromatic (2)	6.8	6.66
7	2.24	2.17	NH2	5.4	5.52
10	8.44	9.00	COOH	12.3	12.33
15	9.94	9.86			
16	2.24	2.28			
***a* = 5.12 ppm**	***b* = 0.86**	***RMSE* = 0.21 ppm**	***a* = 4.95 ppm**	***b* = 0.79**	***RMSE* = 0.095 ppm**
Phenobarbital			**Thymol**		
Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)	Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)
1	10.3	10.49	1	5.4	5.80
3	8.1	8.34	2	6.19	5.90
7a	2.7	2.69	3	7.08	6.35
7b	1.7	1.63	4	3.38	2.91
8a-c	0.6	0.78	5-7	1.05	0.44
9-14	6.9	6.60	8-10	1.45	1.14
			11-13	0.42	1.68
			14	9.99	10.08
***a* = 5.08 ppm**	***b* = 0.78**	***RMSE* = 0.33 ppm**	***a* = 4.93 ppm**	***b* = 0.85**	***RMSE* = 0.72 ppm**
Terbutaline hemisulfate					
Atom Label	Experimental ^{1}H δ (ppm)	ShiftML ^{1}H δ (ppm)			
1	6.83	7.60			
3	6.83	6.50			
4	10.93	10.07			
5	6.83	6.96			
7	4.73	5.26			
10-12	1.33	1.25			
13	7.6	8.22			
***a* = 5.25 ppm**	***b* = 1.00**	***RMSE* = 0.44 ppm**			

Table S8. Experimental and calculated chemical shifts of the structures used in the ShiftML benchmarking. The labelling scheme is given in **Figure S14**. When more than one atom corresponds to a single chemical shift value, their values were averaged.

XIV. Ampicillin lattice parameters

A comparison between lattice parameters of the ampicillin crystal structure, as primitive cell, determined with XRD[22] and NMRX are given in **Table S9.**

	XRD[22]	NMRX	deviation (%)
a [Å]	12.4	11.7	-5.6
b [Å]	6.2	5.78	-6.8
c [Å]	12.0	12.63	+5.25
α	90.0	90.0	0.0
β	114.5	114.506	<0.1
γ	90.0	90.0	0.0
A [Å^3]	839.494	777.213	-7.4

Table S9 Comparison between ampicillin lattice parameter of the crystal structures, as primitive cell, determined with XRD[22] and NMRX.

XV. Positional error estimation

The positional error estimation, using DFT calculated chemical shifts, is done following the procedure described by Hofstetter et al.[43] First, we generate an ensemble of slightly perturbed crystal structures using a set of molecular dynamics (MD) simulations at finite temperatures. By "slightly perturbed" we refer to structures that remain within the same local minima and do not undergo any significant conformational shifts. The MD simulations are done at the DFT level using the universal force engine i-PI[44] together with the Quantum ESPRESSO suite.[24] During the MD simulations the crystal structures were kept at a constant temperature using the NVT ensemble and a GLE thermostat.[45] The used temperatures are given as 1° K, 5° to 50° K in steps of 5° K and 60° to 240° K in steps of 10° K. For each temperature a MD simulation was run during 20 ps and with 1 step per fs. From each temperature we then extract 10 structures at random (5 from the first 10 ps and 5 from the last 10 ps), leading to 300 structures in total with a maximal positional displacement of 1.75 Å, for which the ^{1}H chemical shifts are calculated. This leads to a maximal chemical shift RMSD for ^{1}H of 1.99 ppm. **Figure S17** shows the correlation between positional deviations and the ^{1}H chemical shift RMSD.

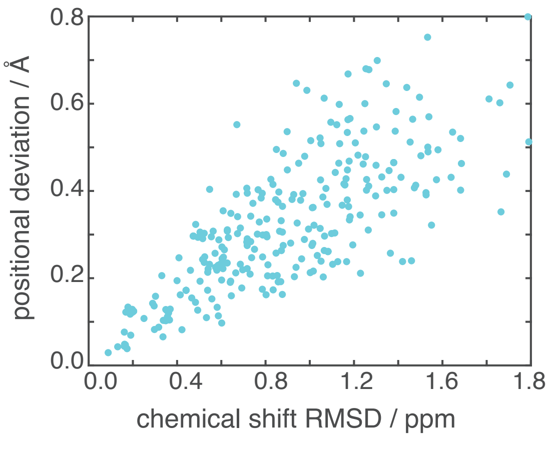

Figure S17 Correlation between positional RMSD (Å) and ^{1}H chemical shift RMSD (ppm) for an ensembles of perturbed crystal structures of ampicillin generated by MD. With $< r_{av} > = \sqrt{\frac{1}{N}\sum_{i,l}\Sigma_{i,l}^2} < \delta > = \bar{\Sigma} < \delta >$, we find a slope $(\bar{\Sigma} = 0.36)$ for the crystal structure of ampicillin.

For the MD DFT calculations the generalized-gradient-approximation (GGA) density functional PBE[25] was used. We used the ultrasoft pseudopotentials with GIPAW[26-27] reconstruction, C.pbe-n-kjpaw_psl.1.0.0.UPF, N.pbe-n-kjpaw_psl.1.0.0.UPF, H.pbe-kjpaw_psl.1.0.0.UPF, O.pbe-nl-kjpaw_psl.1.0.0.UPF and S.pbe-nl-kjpaw_psl.1.0.0.UPF from the pslibrary database [https://dalcorso.github.io/pslibrary/]http://theossrv1.epfl.ch/Main/Pseudopotentials. A wave-function energy cut-off of 60 Ry, a charge density energy cut-off of 240 Ry and no *k*-points. The electron density self-consistency convergence threshold was set to 10^{-8} Ry. For the GIPAW DFT calculations the same parametrization as for the chemical shift calculations of the trial crystal structures was used.

Starting from the DFT calculated chemical shifts of the 300 slightly perturbed structures a continuous correlation function is obtained by maximizing the log-likelihood between the correlation points and a Gaussian distribution:

$$G(\langle r_{i,l}\rangle,\langle\delta_l\rangle) = \frac{1}{\sqrt{2\pi\Sigma_{i,l}^2\,\langle\delta\rangle^2}}\exp\left\{-\frac{(\langle r_{i,l}\rangle - \mu_{i,l}\,\langle\delta\rangle)^2}{2\Sigma_{i,l}^2\,\langle\delta\rangle^2}\right\}$$

where $< r >$ denotes the positional deviation, $< \delta >$ the chemical shift RMSE, Σ the scaling of the variance and μ the scaling of the mean. The indices *l* and *i* denote the atom and the principle axis respectively (PAS). The fit parameters are Σ and μ. The principal values of the anisotropic displacement parameters (ADP) in the PAS are calculated as the mean-square displacements, which for Gaussian distributions is given as the variance, as a function of the chemical shift RMSE,

$$U_{ii,l}^{PAS} = \Sigma_{i,l}^2\langle\delta\rangle^2$$

The amplitudes of the second rank tensors describing the ellipsoids at a given probability (*W*) are calculated in the PAS, where they are diagonal, as,

$$T_{ii,l}^{PAS} = p_{i,l}(W,\langle\delta\rangle)^2$$

where $p_{i,l}(W,< \delta >)$ denotes the *W*th percentile of the fitted Gaussian for a chemical shift RMSD $< \delta >$. These are the quantities that are usually plotted in so-called ORTEP plots as anisotropic displacement ellipsoids, and this is what is shown in **Figure 7c-d**. Note that for simplicity, or for cases with insignificant anisotropy in the displacements, the second rank ADP can be replaced by the equivalent isotropic displacement parameter.

$$U_{eq}^l = \frac{1}{3}\,(U_{11,l}^{PAS} + U_{22,l}^{PAS} + U_{33,l}^{PAS})$$

Note also that from the equivalent isotropic displacement parameters, we can derive a global measurement of the positional uncertainty (U_{eq})for the whole structure, which is given as the average of the equivalent isotropic displacement parameters over all the *N* atoms in the structure,

$$U_{eq} = \frac{1}{N}\sum_{l=1}^{N} U_{eq}^l.$$

The average positional RMSE <r$_{av}$> for a given chemical shift RMSE <δ> is then calculated as ,

$$< r_{av} > = \sqrt{3U_{eq}}.$$

The factor $\sqrt{3}$ results from the fact that the isotropic displacement parameter is given as above, while the RMSE is calculated as $< \mathrm{r} > = \sqrt{\Delta\mathrm{x}^2 + \Delta\mathrm{y}^2 + \Delta\mathrm{z}^2}$.

In addition, the CIF files containing the ADPs without any modification and at the 90th percentile for a ^{1}H chemical shift RMSE of 0.49 ppm are given as Uaniso.cif and Taniso_0.90.cif. The average positional displacements are given for all atoms at the 90th percentile in the file Rav.cif

XVI. ^{1}H-^{13}C HETCOR spectra

All extracted spectral data is given as separate files.

XVII. CSD-3k+S546

The CSD-3k+S546 set contains the additional 1,000 structures containing H,C,N and O atoms with the following CSD Refcodes (in order of FPS selection):

LAWHUM, LASPRT, BANFAY, LITRAH, WATFED, OGEFAI, BOGRUL, UXUYUJ, LGLUAC03, TAHGUE, ZIXPAZ, CEKBAX, TXBNON, RIZHAK, QOXMOG, ZAKKAX, TOVTED, GUBARB10, FEXGUL, JOCKOC, ZEDFIZ, PERLAA, HABMED, WILCOJ, XUYZIC, RIYYAA, PUKGIN, EXIGUO, WANSAG02, DUSMOU, RAJGUG, TOJWIA, TEWHOS, GEDYAR, VESJUZ, BOSPUU, WOTRII, DEXREE, REFROJ, ZAPGII, JEVSIN, QEMGEV, YUYGUU, DHURAC10, CIWMOL, PINGAW, FINTAZ03, OHUXEV, HUGFUL, COKJER, AMIXPO, REWHIL, JUVMIW01, SEQVAN, FIWYUF, KIBFEG, UPONEU, QACXOJ, HAMJAH, UQABUL, ADEWUC, PYRDMD, AFESUA, YOTNEB, LIBWUQ, DIDXOE, DESJOA, KOFKEV, PEGKOC, AIPMLM, YANWAN, IKABIG, EXESEF, QALWOR, JEMTAW, VIPYEZ, FUWPUK, EPAHEJ, AJAROT, MUYTAC, KUZRAZ, WADCUA, IVEZOZ, SNQUOX02, XUDSEW, BZOYAC01, KUPCEF, KOBXUU, NAVQAE, KAHJEI, MAPYNM, NIVPAJ01, VOFZIB, THYGLY10, MOXCIL, ZUJDEP, QOLVAP, ISPADN, JULRIR01, NETROT, KUQVOI, KUKJEH, QIJMUR, QEHDEO, DXTCDA, COKBAF, FAHYOC, DIQCIR, DUDNAR, TEZCOR, DOVCEX, BUGPEZ, EWOVOB, SENRUZ, UFAWUV, IKIYUY, YUQDEU, AJAJEA, JIGJUE10, KUDWEN, WUXKEF, JIYWET, SAWMUA, XADVIH, ZOGPAJ, LIFNOE, TOHWIW, SAWMIO, HODLES, RUNJIV, YAWXEZ, JERHIX, FICYUM, JOHJIB02, BABRIF, GODPUK, CERQIA, ACNPTH10, HALJUZ, SEZVAV, EWAJET, EWUVEY, EZIYES, HEJFEJ, JARRAW, OHEPUO, NUZHUL, JETROQ, KAYTUZ, MIPVOX, CAZDIS, TACQAP, KIDNER, GEPRUQ, MEQVAG, DONTAB, BEKBOI, BUFGAK, ERIXAF, YELWES, GULHAW, DXSDMO, AJAPIL06, FOWVIX, JOHWOU, UJOLUB, GIKTUP, LUMKIP, TUNYUY, BONKOE, EPANUF, LUVHOB, FUYJIS, SERMUZ, MUHRAI, NAXCIY01, RABQUC, JABCEX, IMEGIR04, VAWJEK, HUKRUA, YEFDOE, OROTAC, NOVDIN, AKIKEM, KOMYEQ01, CUKRUW, NUQTOJ, SOJMIN, ISAJUH, UPAQIM, RUJNOA, VIKYOF, JOJRAB, TIMWAN, MENOGU, KOFHUJ, PUHJAE, YIPCOR, ECUXOQ, TEYPIW01, JEWCUK, FACREG, FEFLEI, OVIFUV, MARQIH, DIHTET, RAFWAX, ANISIC04, YILLAG, VOYPAA, QIPFIG, ENANOL, OCIDUZ, XUDSIA, CUJMOJ, LENYUZ, DIXYUE01, DOLSIG10, LOSLAI, WAGMOH01, OLIROQ, LABROV, HISTAN, VENREN, VASTEQ, IFOYAF, JUXJET, BUGDEM, IBPRAC04, AEBHXM10, XAGDOY, IPINEA, EWADOX, TEVHEH, TUMZAE, MCVXDZ, OKIRUX, RUPMEW, ROZPUT, IZUGOZ, PICLIX, XOPWEF03, YUYXOF, MOSKIP, BIMSAR, ODENOB, PUKDUV01, CELGEF, KONHIE, GAKXAT, BEWNAT01, FOHHUF, DMNPYO01, KOYLUH, PYEIDO, UFEPAX, PYRGAL03, KOBXOO, IJETOG02, LALQEU, NUFRUB, LEFRAP, EJALOR, XOMBEH, NIHFAN, UPUKOH, AJOFUA01, ZOXCIA, HASLIY, POVMAQ, GAGVAL01, KAVQON, HMBZQU, LIKDIS, GAZGOF, NUYQEE, JAKSOD, BISMEV11, PCBUTO, CUTTOB, BEHBOG, NOPQIT, DATQAR, FAWNAS10, NIPPEJ, IFIYIF, MIHTEE, GEDYEU, QATTIO, SALYEL, NIGZEJ01, BATWUO, FEHVEU, ROGVIU, UFUXOI, NPYRAM, NAKJIT, PIPNFP, ZIYZOY, WOLNOC, YICNAA, ODOHIA, FIBGAA, WARCEZ, ZZZFKK01, VIDLIE, VOSDIQ, UNUKUK, MULPEO, PAHYON01, USIWEZ, PUGTIV, EPEZOP, EVOZIZ, TELZOZ, ADUWUS01, VAYLAI, VIBCEQ, YUKVAB, METHYM01, VOZHUN, GEXQEF, MCBURL10, DAVYAA, IRUQIX, QOGPOR, VAVHEH, INACOQ, PAXQUC, UJULIU, VAGTED, MELWUW, PECWEA, NADYUM, KEPKAR, GUGMIG, QETROW, FOYGOR, YACLOD, ROSMOD, GLGLYN10, MOGYIR, JAHZIB10, UKIJEF, FINFUE, BAGFIY01, IFETAO, ODEZOO01, CIVVEI10, FADYIV01, OCETUN, MEADEN05, SAJYEH, GACSUB, UTALIM, YUQBAO, ZZZGWW02, PAXTIT, QIQBAV, RSAMPA, DAFWAI, PYRGLU10, SOYVEJ, WOGFAZ, BUHYEI, JEPPEZ, BUBXAY, FAYDOZ, VIHTUD, MXBPOX, NOCPAY, DANJEH, RARLIG, FAYJAR, SUNJIU, VEKQEH, DUMPAC, TETZOL02, OQOSOD, XUMYEL, AFIMOT, MEGWUR, KORXAS, QOZFER, VIWCOT, KOWYEA12, WABSIE, WEPWAQ, GLYGLY17, ZORJOF, PAZFUS, CORRUW, HEJGOU01, ULIDID, FORMAM, CAVQUM, PYRZIN19, KEQZOW, VOHXIB, CEPROF, OXOFMB, UVEGIN, WAKFEU, UHODIG, WEPTUH, DAMTRZ11, CUPYOD, PELKEY, PATXAJ, FARTID, TUQDAM, MELAMI04, GLYCIN16, BOKSIE, EROFEX, DIHPEQ, AMHMAN, SEZQIY, VIKKEG, XUHZAD, FUMYOD, RAPRUW, LOTYEY, PEFGEO01, OGEGIR, HIVTUD, TATYAQ, LIQPEH, PECWAX, JAKKEL, KOVGOR, SADXAW, VUHSOI, HEHXOJ, FURALA10, PRMDIN01, FOCZUT, BOLLIY, LUDFAR, VAVVET, CEWGUH, CTHXDL, CIGJAD, NUZSIM, GEFNAH, MPYAZO10, NEYMOT, XOJHOU, ONIVOY, XAJSUX, DUCYOQ, FOMJEX, VICREE, CUKGIZ, ZEHREJ, IPUMAH, DIALAC02, AMSALA02, HOBXEC, PUKMOY, CEDKON, LICLEP01, YAHYUD, PEVVER, LONKIK, NUMHIN, SICSEC, NULCII, QOBGUL, WOMYEC, SAPXEM, NIXPUF, FIFFIK, PEMQUT, CAKXAO, KAXWOW, PAZDPY, VUQZOX, KIYQUF, FUSFAB, BIDCEW, FETXIM, CAZBUC, DLNLUA05, JELSEY, WOBLUU, TILJUV, LUDFOF, QUMWEB, SOVNEX, RUWFIY, HODKAN, IKAQAN, NUPBIL, NICOAM06, NISMIL, KOFZOW, ACETSC10, IQOMIM13, RAYXUK, NMHXNT, SAHPOI, TUJKIT, DAPJUA, SATRIO, VINXEW, AHANOM, LINKAW, ECELED, PURINE01, FAVMIY, JUKJUW, ENIJEQ, ZEGCAR, YETLUE, LUJMEK, YAQLOS, XUCJEM, RAHSEY, PAMQAW, PERXOA, KOJSUZ01, BEHBIA, MAJTEX, NIQPOT02, MEQPED, PEYZOK, GIJLIW, BACXAE10, SOLVEV, WIJNEI, XAYVIC, PABHAB, TIHJEA, GUACET, OTUCOW, QADBAY, ULAVOU, XUTJIH, XOSCAK, VANXEO02, NEHPUN, YAJXOX, JECCOJ, JESJIB, BILXUR, HALMAK, KEKCEJ, INAFIN, YIDLEC01, WAKROS, VAMJOJ, KEDWEV, ZOYCOH, XELJUT, KINREE, VACDEK, UTETOE, FEMGUZ, PUJZEA, PAZCUO, FOXWAR01, MALROH, ZODWIY, HUFSUX, QIZZUW, SOLLOV02, UWAKEK, FIQTIK, WIBCAL, DOLLEV, RALCEN, PUYROQ, JUBCEO, CIWVAH01, LINJAT, XEHHAU, UBEMIX, EFOZAB04, ACERUX, DZCDON, BOQVIM01, MODYUA, FACVUB, SARBUI, GUXFIQ, VUPTAC03, LEPCIU, TIBMUM, KUTHEN, PAHZOO, TEDQOK, RIYQOG, GEJBUU, XIYXUZ, SESQIR, BARKAH, ENOBEP, KEBGAB02, BOMKUJ, ITIWIT, TAXSOA, BOQWUZ, DAVXAZ, KABMIL, OLORUC, VEVVOJ, NEFLOA, MOBMEV, KIQZOZ, KUXSIG, JULZAR01, REZMOY, ESOZAP, OLICUH, PAJNIX, BUGKOE, QIXXUS, QAGQAT, CBCPIC, YOTYOW, FORSAI, YOKJOY, PHYPHM, ACOWIY, ZIBBET, NBZOAO06, DIWPOO, DEFYAO, GLYCIN02, GOCJIS, BIHKEI01, HEGFOP, UHACIQ, VUWKII, HIVLAA, CAZBOW, EKEPEP, FIZDIB, HALMAJ, CIVTIM, HEYRIM, ISEPEB01, FEBBIZ, JEHGEI, EXOQOY, COWGEB, YATGIL, DOBPIV, PUKFAD, YOVPAB, PRFUXC, TESKUZ, TAJHOB, KASZEK, TAWPEN, OKAVAY, TERSEP, AXASUO, BOLFUD, UTUSUZ, LEJKOA, PACRES, CIWWAI, FEXBUF, XUJPAT, SOWWIK, VEFXOV, WUDJOU, LEKZEH, SADBEE, NENXAF, BEGDIB, IKANUE, DIXNUT, ERISIH02, WAFGER, XUSROU, NOPSOA, LAYYUH, QAHPIZ, JEMRIC, WABNOF, AHICIF01, LEFJEM, XUVXAN, AYAWIH, TBFRAC01, BEPTUN, UXAZID, NORGEH, EPAHIN, DOGPAQ, IVALUO, ZEJJED, NIFDAI, NEVYUI, HAMXEA, ILOHIC, VACROG, UPIKUB, XARBAV, UNOKIR, MESCON, ZADKAT, GISDIX, OXUMAW, DOLTAZ, TEHMUO, TALTIK, BETBEH10, PABNAK, FOMTAE, AMPTRA10, WOCHED, BOTMUT, ANITOI, TABDUY, QACZEB, GOKFUJ, HOQZEU, UKIKIK, XERPIU, BIGSOZ, XAVMAJ, LIMLOI, XAQNOT, ZIBBAP, AHOXLH02, LIDCAE01, HUKGOJ, OPUSIC, OTUDIR01, SADCUV, UXUKII, EBIROX, LOWBIK, TOLCAY02, URACIL, DUTCAX, TAWPAJ, AMBNAC04, MIMMEB, QAKCOU, ICYTIN, POFKAY, CEKTOB, EVOPOV, GUZYEF, DOGQAR, FUHTAE, BEVLOD, VELNOR, CAGVIQ, TIYLEU, ZIKTIW, GLYCIN62, XIFRUA, VICBAK, JESMEA, GOBPUK, WUNCOZ, JUFKIE, OJOCIB, JAGMEM, NADREN, NEPNTH, FORMAM03, TAVXOF, PATNUU, NERJUQ, VABJOX10, XERBEB, GINMOG, KUGVEP, EVATUR, EMUMOQ, LIPVUB02, VIWYAD, PARGIY, POXWEE, MNTBZA, PHBZAC, TAYGUW, IFUPUV, BOBCUO, NOSLEM, LIMKIB, EFOPEW, ALLCAM, UTIXAX02, DOKQIE, QIKZAN02, HYQUIN, ACOHAC, WUXYAQ, MECPYZ, REZNUG, XEZBAG, BEPTUL, POFNOO, PUWFET, JUPVAR, BEMCAZ, CAKRIQ, KAWQEF, CERQOG, WEJHUO, PATTAF, WOFYOH, KAKSEW, DORJEA, RICFIS, QIPPUA, HEWWAI, IWATIL, XIXTUU, RUJQIX, PASZIS, KOKTEL, NAHDOP, OKOTAL, FANRAO, JURQAO, COXZAS04, LONDOH, FOWDAY, FUQLAE, WEHYAK, JODHOA, BUWHUW, IBUVUY, DIVXIP, GITNUT, TUHCEF, DUBRAT, PEMQII, AWELOE01, FEGHAA11, DELXOH, CEXQIG, COWYAP, HIYDEA, PEYZUQ, JEYNOR, JAKCOO, VABZUV, CIHWUL, CAXMOD, ZAJGUM, JIWZAQ, RAQROR, PECZOO, ZOWDOE, EXIMAA, ARUBIY, MESUCC, EXIVUD, OMEVIL, XUKDEN01, ARAQUH, CADHAS, CARCAZ, MOSCON, LUKNAG, IXINAF, HASHAK, FUGJUM01, MPPAZD, CHONCL, SOYRII, JUMXIA, GLURAC06, LUGJII, FUSJAF, ZIKKUZ, UHUCUW, AXACEI01, UHIQOS, LIXGOO, MOSJOU, ULICOJ, ILIVAC, MEXCUP, IFAYEV, JEJZAA, SUJRAS, SIGGUL, CIKGAE, VOQMOE, JOWSOD, SOFTEM, CUPZIY, TISVEX, LAQSOM, DUWRIW, DUPHOL, QOHCIA, BASBON, PEMSOQ, LAPDIR, LETBUH, TORTEZ, GIXFID, MOKYER01, HIRRIK, BCBDNT, FUGQED, GIMREA, PROLON01, GOKTAC, YABGEP, ABUBEE, FADSAF, PICAMD06, BCOCDC, NESTIP, CIKXOJ, AHIHEF, GEWGUK, BASXUP, HMUSCM, ZIPWOM, NTBZAM11, KIXCUP, AZTCDO10, XEWNAP, LIPSUA, QEPFIB, VIZCEO, EXUQAP, EJISUN, NUHLIN, OBUSIN, WATDAW, AGAWIQ, VEYVOK, IJEKAJ, GUVBAC, JEMJUG, FOGWED, NIJPEB, RAMVOQ01, YOPKOE, COYNAF, YOVTUZ, MAMVUS, CACHAP, XIBTEJ, WAVXIA, SUPJUI, CIWNED, BONGAO, LUPWOI, DEHSAK10, MAYWIU, HNIMOZ, NUHYOE, SIHPOQ, FUTDED, NADMUA, KUGZER, YIHLEI, LUMJOU, HAJMOU, EMOHIY, LALCOQ, RASDUM, XEJYAN, XODDUQ, PONNIQ, ABIRUZ, OKOZIZ, MENMEY, TARTMM01, FUNVAL, YECWAE, VOBLUV, NUTMOE, MODCUD, KOJZEO, WOLGOV, QUFPEO, OXAREL, BUFCEM, PIWDUU, PIPMDC11, NUZQOO, VIPSAP, DEKYIC, PAXMOS, HUYNAQ, AJEGUS, MINCIW, ANUREH, CIGGAC, XIBSUX, EFITAP, YAMBEV

The CSD-3k+S546 set contains the additional 546 structures containing S, H,C,N and O atoms with the following CSD Refcodes (in order of FPS selection):
DOWQAJ, ACOTHI, AMTHTZ, AXOBAS, AMTCAR03, ELEWUN, BDTOLE02, QARZIV, YIDBAO, HIVVUF, NAHMUE, TAPBAP, YIGJOP, PADQAN, XIWBAI, PESGOJ, VEPSEQ01, ICAKEC, MOLJED, EFOTEY, LERJEX, TSCARB01, IHOZUA, FUWMIT, UYIVAB, KABLUX, CEHQEM, MUQBUV, MADGAC, PEBTAU04, WOCQEK, VOBMEE, DTUREA10, SEFZIM, HOQCOF, FUNDEY, RUYSUZ, SIHZAK, SUPFOY, ISEBUD, SIYBEJ, IHAVEU, AZOTEP, BIRFEN01, HUHWAI, JARDEM, DOCNIS, TANTAD, FIYZAO, RAPGAR, OHUSIV, TBPMTK10, MESCZA, FOGHOY, OJIVOU01, LEFLUE, PTZCNB, LINCIU, KAVYIP, NUXGOC, ZZZVZC01, QOBFUI, HTDZDX10, DAHDAR, KEXNIK, XARHII, XUPGAQ, BEXSAY, ULEWIR, DSCARA, AFIXUK, AZAZEI, MTDZDT, CYSTAC08, IFIZIG09, TUJWOL, SIXXON, VECXIL, WADSOL, CAFVUA, UVALIN01, EROMUU, RELLAV, NEZNEL, SENGAT, RHODHY, NAKWAA, LUYKOF01, PUQHIS, GIPVUW02, GUCDEO, LISTAJ, QIYRIZ, KOYMES01, EZOPEP, MMCPUR, JORLOS01, DEMQIW01, BEBPAA, BUXCEC, PAMJAO, DIRMIA, THCHYD, SUZLAA, PUVKAT01, LIHXUW01, QEPLON, NIQVAL, WAXNUH, JIZFAZ, GAJTEQ, HEMYON, XUBLEN, HAJVOE, DMETSO, SAFGIP, YEFXAJ, VUKZIL01, FABDOE, MACNHS, UJIRUC, FAQWEB, ZATMAK, FOWTER, DIZVOZ, MAYGEZ, SOCGUM, BANSUF, BIXNEB10, AJIGAB, MAPZAF, OMEQUS, SIKJIF, CECQAD, ODEDAC, DAHFOI, VEZCAE, CADCIU, DOMDUE, DAHHUS, FUHZOY, XIGSIR, ZERRET, RAHNOG, GADPAE, ATTDAZ01, XIBZOX, KANVAX, OMAXOQ, GUGJOH, ZOSRUU, KABJUT, IPUCUT, CIGYIA, XOCJEE, QIQCUQ, FEYDUJ, REGYEH, ASOPEF, MACTUH, CABCUD15, RASYOZ, XANHYD, KIHZUY, ZUWPOW, ROGWOB, IVAQIH, UFILUQ01, NUYMUQ, BABFUI, ELIPAR, EFOQOG, DICHEC, WAFLEW, GEMLUH, REHMIA, LIJXOR, KADJIL, TEJPEE, SIYBUZ, HORZAQ, ONOHEF, TIJLAZ, FOWQOY, PUNMAO, LCYSTN25, TCHYZS, HILVOO01, GEHPUE, LEWZIY, JIQZOY01, COKXUX, DTBIUR, IMICOW, KURRIY01, YILWEV, ASODUJ, DUVFUV02, COSLOL, RHODIN01, JATNEX, PAMPAX, ATZTHD10, UQISOE, MUKJIL, MNDXTO, XEZKAO, TARBUL, VUYWUH, PYZPYT01, IGISIA01, SOJNAG, DMETSO06, COWNEJ, CEHJIJ, DTURAC01, FEDNUY, FOWFII, SUFWIB, EHAXOB, DMDNTP, ILABEE, UJOSER01, IXANOK, CUGLUN, OVAYAM, JUPPAL, YECGIZ, KUFDUK, WOWFOD, BIKFAC10, VONWAX, TUPRBN01, FOGJAM, BOQCUF01, XADSOK, IRILEB, SIKQOS, NOJKED, ADPTHZ, FEZRAD, FUSVEW, PETNEI, SURSUT, UCOVUD, KABFID, BUDPIB, DOJPAT, UJIBAR, XEZFIT, NIBXOL, SIGCAM, AVULIM, XAHLOJ, XSHCXB10, POTLEP, ATDZSA04, RESHAY,

BUBNOB, HAHDOL, YIHTUG01, WACJIT, AYAHOY, PORQAQ, CALJAD, ICOMUI01, QILZIU, QACPOA, XIHKUU, LUVYEI, HTHSOC, JEXROT, CERPEV, BASVOG, NINTAF, VEQVAP, XEWKAN, LUHJED, GISKEA, WENVIV, SIDKOG, QETKOP, ZEWDOU, YEWLUI, HIWNAE, JULMIM, NEXSIU, GEHBAX01, GEFTES, GEXKID01, AXOSEN, LEJPAS, YIFGAW01, NINZIV, WIFKOL, NOGKEZ, HETAUR20, LORXOG, COXZEU, THACEM, HOMGOG, CASHOT03, AGLYSL, DODNUF, UQAFEZ, DTHDOM, DIQCAJ, DELZOK, YISBOR, MEHJEQ, ELATUG01, ETTHUR04, FEJYIC, VUCJAF, WINMOV, SUNDOW, XALXAK, UJIWEQ01, AFUTDZ10, ZEBYEM, PUXDOB, FOQKIF, IPAWAX, AYUBUS, NELQUQ, QIPVEQ, KACGIG, DUVYEZ02, ACOJAF, MERPUM, WURJIE, FADBOB, ZEYBAG, OXTTCD, XAZTID, TARKAZ, PAMSUS, AKEKIL, OBUWOA, DTSUCC, BEZQAZ03, RUWRAC, DIVRIJ, QEHSEB, KUQQEU, SOYJAR, KUZWUY, BETBUX, OXUVUZ, JEFMUD, POVSOI, TOMBAA, TURACD, HURXOH, MOSYUO, WAGKUM, MBZOXT, EPEDEK, OCIBIL, MOPSER, WAXCUV, WONFUB, FAGRIO, XOJSAS, SAFYIK, HAVSEC, NSBTOA, KOVHOS, ITAFUF, NUFCOH, ROGQIO, AHUFUF, CECPIK, NEWYIX, QOQVOH05, PAMGIW, PYRIDS03, WETFUX, OZAHEE, DAHPIL, DMTRZT, VUZHEE, TACMSO, VECSAX, ZAKJAW, KOKNOO, OWEKEH, DELMOX, TASPIN, DAPXEX, ULEHAV, XUVXIW, GUFJAS, HESSIH, RABVIC, RAZQOB, DUKWUB, PIWDIJ, YILPAM, VAFZOT, MEMWUW, GAZPIG, WETPIU, TECCIP, MUCCOD, AMPTOP10, BATSUK, DEFYUI, WOJPUI, ZIYZEM, LESJAV, HIXKEF, VISDIK, CAGJEC01, UWOSAC, PEBWOI, KABJED, PENLID, PEMBEO, SUVWAJ, PEJVOQ, KODXEI, PIKGIZ, TOBRUX01, QISQEP, GOQBAP, ZAZYIK, TUHFUX, DOZKOS, MPYDCX, CICTEN, WOKLUF03, HIYNUY, TUBCOI, BOZKOQ, ODAGIM, NITPUB, SOVZAE, DAMDAZ, TRZTHO, BOZKUW, ITUBEF, HEQZEK, DTCZME, GAQZOP, AGEVOZ, XEDDAM, THPYDO, ILOLOL, XEHDAS, ODIKES, SOCZUF, GESSAZ, LULLAG, JOTSES, KAYDIZ, PYMDSD, WATKEH01, LANXON, UZOJIE, UMIXIZ, ACAYIM, QUBRAH, THIOUR14, VONZIJ, LANSOK, DMSCPY, TAZOLD01, HOQGEB, RONVEV, WEGSIL, MARFIV, JIGCIL, LORQUF, EVOQEL, DOHGOW01, TAURIN04, YEPXAS, YOTSOR, HIVWUF, AJUXEI, QEGBUZ, UMIQEO, WEWPUJ, FOYNEO01, GUMCOI, HMDTOX, GAVNUO, COQQEE, DUJHEV, CEPNAN, MORPDS01, KAKQAO, DEHYUM, GUWPIZ, HAXPAX, DAXSAZ, AJIGEF, DEDWER, TOSQUN, KIVDAV, EZOMEM, MAKNIX, VUKRUP, UKIKEF, XOJKAK, ECAGIA, QEKXUZ, NOYWED, WEWRAR, NOFYIQ, TAJWAC, SUNDIQ, ANOTAZ, ABAROL, DTHKET, NEZNIP, ULUTEA, ADOFEF, KADMIO, IZAWUB, AGEREJ, HINNEZ, GAKPEN, AYAYOP, GIZBUN, BUKVAE, KAGMIO, QENYOX, GISYEO, PEBCOO, FIZHAX, HIQQED, LUXPAW, CANCIE, DUZFUB, REWYEX, AHOMAM, FEWMAX, WOLVUP, BAHPAC, FEMGIN, PEVSIS, USUWOW, KEFZUR, ETHUSS03

XVIII. CSD-500+S104

The CSD-500+S104 set contains the additional 104 structures containing S, H,C,N and O atoms with the following CSD Refcodes:
KURZOO, DAGMEF, EZUMUJ, NILSAE01, AVULUY, HOTNUA, AKOFUD, UCAYUS, LAWCIW, VEKHAX, CUSTAN, KELNIA06, RAMJIB, ADITAI, SEYMEO, VUJHOX, ETABAC, POWBUZ, ZOYPIM, LANQIC, GUFVIN, TIPLEK, KOXWIF, RIYCUZ, MAXVUF, YUPXUD, KUSVOK, DUGGET, NAPKAR, TUJPEV, GAQHIR01, OFAQIV, TAPHUP, TNONDX, WIWZAF, ZAPJUX, CEPSIZ, COMLOH, OSOJOW, OJUYAV, ZUYJEK, PUGNUB, ZZZPUS09, AJOGUC, ANONIB01, WIJDIE, CIDYIZ, LAWTAH, ZEYKAR, PARJUQ, ZOHNER, VAHKIZ, LODWAE, GIKQIA, KISXOB, LAQPOK, TEYDUW, HIRWIP, METZHY, AKEVOC01, PAVHOM, YIHTUG, JEBLEK, UNOPAQ, HULQIO, NINREI, TENWOA, MADTET, VAKJAV, LABVIU, RICXEG, ABEQII, MPAZTP, MACMOT, LEGREW, QOKVUI, PAHDIM, GOLXAH, MIPLUT, VEBPUQ, SUXYOZ, FEPJER, PDTHON, HAGXET, JARDUC, CATBOP, GAVPAX, ALAQEK, CACYIQ, SIDBUD, HIQPEC, MOVBAA, TIHJOK, YOFTIY, TEBFOV, QIFDIT01, PIKGEV01, TIYXII, ISULUF, SODGID, NAJBAB, FOGVIG04, VINNEM, IZALUQ

XIX. References

1. Boles, M. O.; Girven, R. J., The structures of ampicillin: a comparison of the anhydrate and trihydrate forms. *Acta Crystallographica Section B* **1976,** *32* (8), 2279-2284.

2. Cense, J. M.; Agafonov, V.; Ceolin, R.; Ladure, P.; Rodier, N., Crystal and Molecular-Structure Analysis of Flutamide - Bifurcated Helicoidal C-H ... O Hydrogen-Bonds. *Struct Chem* **1994,** *5* (2), 79-84.

3. Hrynchuk, R. J.; Barton, R. J.; Robertson, B. E., The Crystal-Structure of Free Base Cocaine, C17h21no4. *Can J Chem* **1983,** *61* (3), 481-487.

4. Murthy, H. M. K.; Bhat, T. N.; Vijayan, M., Structural Studies of Analgesics and Their Interactions .9. Structure of a New Crystal Form of 2-((3-(Trifluoromethyl)Phenyl)Amino)Benzoic Acid (Flufenamic Acid). *Acta Crystallogr B* **1982,** *38* (Jan), 315-317.

5. Hayashi, S.; Hayamizu, K., Chemical Shift Standards in High-Resolution Solid-State NMR (1) 13C, 29Si, and 1H Nuclei. *Bulletin of the Chemical Society of Japan* **1991,** *64* (2), 685-687.

6. Baias, M.; Widdifield, C. M.; Dumez, J.-N.; Thompson, H. P. G.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M.; Emsley, L., Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy. *Physical Chemistry Chemical Physics* **2013,** *15* (21), 8069-8069.

7. Clayden, N. J.; Dobson, C. M.; Lian, L.-Y.; Twyman, J. M., A solid-state 13C nuclear magnetic resonance study of the conformational states of penicillins. *Journal of the Chemical Society, Perkin Transactions 2* **1986,** (12), 1933-1940.

8. Antzutkin, O. N.; Lee, Y. K.; Levitt, M. H., 13C and15N—Chemical Shift Anisotropy of Ampicillin and Penicillin-V Studied by 2D-PASS and CP/MAS NMR. *Journal of Magnetic Resonance* **1998,** *135* (1), 144-155.
9. Kolossvary, I.; Guida, W. C., Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *Journal of the American Chemical Society* **1996,** *118* (21), 5011-5019.
10. Kolossváry, I.; Guida, W. C., Low-mode conformational search elucidated: Application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. *Journal of Computational Chemistry* **1999,** *20* (15), 1671-1684.
11. *MacroModel*, V9.0; Schrödinger LLC: New York, NY, 2011.
12. Harder, E.; Damm, W.; Maple, J.; Wu, C. J.; Reboul, M.; Xiang, J. Y.; Wang, L. L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016,** *12* (1), 281-296.
13. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *Journal of Computational Chemistry* **2011,** *32* (7), 1456-1465.
14. M. Cerrioti; S. De; R. H. Meissner; Tribello, G. A. Sketch map package. https://github.com/cosmo-epfl/sketchmap/.
15. Ceriotti, M.; Tribello, G. A.; Parrinello, M., From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011,** *108* (32), 13023-13028.
16. De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M., Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016,** *18* (20), 13754-13769.
17. De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M., Mapping and classifying molecules from a high-throughput structural database. *JOURNAL OF CHEMINFORMATICS* **2017,** *9*.
18. Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016,** *12* (2), 910-924.
19. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
20. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.
21. Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C., First principles methods using CASTEP. *Z Kristallogr* **2005,** *220* (5-6), 567-570.
22. Boles, M. O.; Girven, R. J., The Structures of Ampicillin: a Comparison of the Anhydrate and Trihydrate Forms. *Acta Crystallographica* **1976,** *B* (32), 2279-2284.
23. Charpentier, T., The PAW/GIPAW approach for computing NMR parameters: A new dimension added to NMR study of solids. *Solid State Nuclear Magnetic Resonance* **2011,** *40* (1), 1-20.
24. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter* **2009,** *21* (39), 395502-395502.
25. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996,** *77* (18), 3865.
26. Pickard, C. J.; Mauri, F., All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys Rev B* **2001,** *63* (24).
27. Yates, J. R.; Pickard, C. J.; Mauri, F., Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys. Rev. B* **2007,** *76* (2), 024401.
28. Monkhorst, H. J.; Pack, J. D., Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976,** *13* (12), 5188.
29. Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L., Chemical shifts in molecular solids by machine learning. *Nat Commun* **2018,** *9* (1), 4501.
30. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallogr B* **2016,** *72*, 171-179.
31. Bartok, A. P.; Kondor, R.; Csanyi, G., On representing chemical environments. *Physical Review B* **2013,** *87* (18).
32. Willatt, M. J.; Musil, F.; Ceriotti, M., Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics* **2018,** *20* (47), 29661-29668.
33. Carignani, E.; Borsacchi, S.; Bradley, J. P.; Brown, S. P.; Geppi, M., Strong intermolecular ring current influence on 1H chemical shifts in two crystalline forms of naproxen: A combined solid-state NMR and DFT study. *Journal of Physical Chemistry C* **2013,** *117* (34), 17731-17740.
34. Uldry, A.-C.; Griffin, J. M.; Yates, J. R.; Pérez-Torralba, M.; Santa María, M. D.; Webber, A. L.; Beaumont, M. L.; Samoson, A.; Claramunt, R. M.; Pickard, C. J., Quantifying weak hydrogen bonding in uracil and 4-Cyano-4 ‘-ethynylbiphenyl: a combined computational and experimental investigation of NMR chemical shifts in the solid state. *J. Am. Chem. Soc.* **2008,** *130* (3), 945-954.
35. Sardo, M.; Santos, S. M.; Babaryk, A. A.; López, C.; Alkorta, I.; Elguero, J.; Claramunt, R. M.; Mafra, L., Diazole-based powdered cocrystal featuring a helical hydrogen-bonded network: Structure determination from PXRD, solid-state NMR and computer modeling. *Solid State Nucl Mag* **2015,** *65*, 49-63.
36. Harris, R. K.; Jackson, P., High-Resolution H-1 and C-13 Nmr of Solid 2-Aminobenzoic Acid. *J Phys Chem Solids* **1987,** *48* (9), 813-818.
37. Hartman, J. D.; Kudla, R. A.; Day, G. M.; Mueller, L. J.; Beran, G. J., Benchmark fragment-based (1)H, (13)C, (15)N and (17)O chemical shift predictions in molecular crystals. *Phys Chem Chem Phys* **2016,** *18* (31), 21686-709.

38. Tatton, A. S.; Pham, T. N.; Vogt, F. G.; Iuga, D.; Edwards, A. J.; Brown, S. P., Probing intermolecular interactions and nitrogen protonation in pharmaceuticals by novel N-15-edited and 2D N-14-H-1 solid-state NMR. *Crystengcomm* **2012,** *14* (8), 2654-2659.

39. Abraham, A.; Apperley, D. C.; Gelbrich, T.; Harris, R. K.; Griesser, U. J., NMR crystallography - Three polymorphs of phenobarbital. *Can J Chem* **2011,** *89* (7), 770-778.

40. Salager, E.; Day, G. M.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L., Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution H-1 Solid-State NMR Spectroscopy. *Journal of the American Chemical Society* **2010,** *132* (8), 2564-+.

41. Harris, R. K.; Hodgkinson, P.; Zorin, V.; Dumez, J. N.; Elena-Herrmann, B.; Emsley, L.; Salager, E.; Stein, R. S., Computation and NMR crystallography of terbutaline sulfate. *Magnetic Resonance in Chemistry* **2010,** *48*, S103-S112.

42. Baias, M.; Widdifield, C. M.; Dumez, J. N.; Thompson, H. P.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M.; Emsley, L., Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy. *Phys Chem Chem Phys* **2013,** *15* (21), 8069-80.

43. Hofstetter, A.; Emsley, L., Positional Variance in NMR Crystallography. *J Am Chem Soc* **2017,** *139* (7), 2573-2576.

44. Ceriotti, M.; More, J.; Manolopoulos, D. E., i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Computer Physics Communications* **2014,** *185* (3), 1019-1026.

45. Ceriotti, M.; Bussi, G.; Parrinello, M., Colored-Noise Thermostats à la Carte. *Journal of Chemical Theory and Computation* **2010,** *6* (4), 1170-1180.