A Bayesian approach to NMR crystal structure determination
Supplementary Information

Edgar A. Engel,[1] Andrea Anelli,[1] Albert Hofstetter,[2] Federico Paruzzo,[2] Lyndon Emsley,[2] and Michele Ceriotti[1]

[1] *Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

[2] *Laboratory of Magnetic Resonance, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

(Dated: October 4, 2019)

I. A BAYESIAN APPROACH TO STRUCTURE MATCHING

A pragmatic alternative to the Bayesian approach is to approximate

$$p\left(\mathbf{y}^\star|M\right) \approx p\left(\mathbf{y}|M,\mathbf{a}^\star(M)\right) = \max_{\mathbf{a}} p\left(\mathbf{y}|M,\mathbf{a}(M)\right), \quad (1)$$

assuming that the sum in Eq. (4) of the main text is dominated by the contribution from the "best match" assignment $\mathbf{a}^\star(M)$, which can be computed easily using the Hungarian algorithm [1]. This approximation is consistent with the conventional strategy of computing the RMSE between experimental and predicted shifts based on the assignment that minimizes the error. In the main text we show that it captures qualitatively the relative posterior probabilities of different structures, but is not quantitative, and can fail to determine the most probable assignment when there is a subtle competition between models.

II. ASSIGNMENTS OF ^{1}H AND ^{13}C CHEMICAL SHIFTS IN MOLECULAR SOLIDS

While there isn't a unique way of performing chemical shift assignments, we here illustrate a typical protocol for assignment of ^{1}H and ^{13}C chemical shifts for molecular solids at natural isotopic abundance. Starting from the chemical formula, the first step involves acquiring the 1D ^{1}H and ^{13}C solid-state magic angle spinning (MAS) NMR spectra. Given the large line broadening typical for ^{1}H solid-state NMR, the ^{1}H spectrum should be acquired either at fast MAS (above 65 kHz) or by combining lower MAS rates with homonuclear dipolar decoupling techniques (the so-called CRAMPS experiments) in order to maximize the resolution [2–4].

The ^{13}C NMR spectrum should be acquired using cross-polarization (CP) to transfer the magnetization from ^{1}H (which have higher sensitivity) to ^{13}C. To distinguish quaternary carbons, it is possible to acquire an additional ^{13}C spectrum with short CP time, where only the signals of the ^{13}C directly attached to ^{1}H are visible. If nitrogen is present in the molecule, it is also useful to acquire the 1D ^{15}N spectrum. The resonances in the ^{1}H and ^{13}C spectra can then be tentatively assigned using theoretically predicted GIPAW DFT or ML [5–7] NMR chemical shifts. For complicated molecules, comparison with solution-state NMR spectra can also be used at this stage.

Next, this initial assignment should be refined, which is most easily done starting with the ^{13}C spectrum. 2D ^{13}C–^{13}C INADEQUATE experiments [8] provide correlations arising from neighboring carbons, which usually suffice to unambiguously identify all ^{13}C resonances. These experiments typically have very low sensitivity, and so they are often paired with dynamic nuclear polarization (DNP) to facilitate the identifications [9].

When the ^{13}C spectrum is fully assigned, it is possible to assign the ^{1}H spectrum. This is most simply done through ^{1}H–^{13}C HETCOR experiments with very short CP contact time. These provide ^{1}H–^{13}C correlations only between the directly bonded ^{1}H–^{13}C pairs, allowing the assignment of the ^{1}H resonances based on the ^{13}C assignment. Alternatively, J-based ^{1}H–^{13}C correlation experiments such as MAS-J-HMQC or the MAS-J-HSQC can be used for the same purpose [10, 11].

If multiple hydrogen atoms are attached to heteronuclei in the sample (such as nitrogen), it is possible to further perform ^{1}H–^{15}N HETCOR experiments with short contact time in order to understand the ^{1}H–^{15}N connectivity.

III. APPLICATIONS

A. Crystal structure prediction

Detailed descriptions of the generation and refinement of the candidate crystal structures for all compounds discussed in this work can be found in the original publications [12–15]. In summary, the theophylline, flutamide, flufenamic acid, cocaine, and AZD8329 candidates were generated starting from their chemical formulae using CrystalPredictor [16] (and, in the case of ampicillin, the Global Lattice Energy Explorer code [17]) to perform a quasi-random sampling of unit cells and molecular positions within the most commonly observed Söhnke space groups, all with one molecule (geometry optimized using DFT with the hybrid B3LYP functional [18, 19]) in the asymmetric unit cell. For cocaine this was prefaced by an automated conformer search using the low-mode search method [20] leading to 16 starting conformations, while for the other compounds a search of their torsional

energy profiles [21] provided eight (flutamide), six (flufenamic acid and AZD8329), and 16 (ampicillin) starting conformations, respectively.

Subsequently, the **theophylline** candidates were geometry optimised at fixed molecular geometry using the DMACRYS code [22] with the FIT potential of Coombes *et al.* [23] and electrostatics based on atomic multipoles from a distributed multipole analysis [24] of the electron density at the B3LYP/6-31G(d,p) DFT level of theory. For **flufenamic acid and flutamide** the candidates were geometry optimized using a molecular mechanics description of inter- and intra-molecular interactions using an atom-atom model with exp-6 + atomic multipoles electrostatics and B3LYP/6-31G(d,p) DFT, respectively. The influence of polarisation effects was approximated by performing the molecular calculations in a continuum dielectric ($\varepsilon = 3$). For **cocaine** the lowest energy structures were geometry optimized using CrystalOptimizer[25] using the same description of the intra- and inter-molecular interactions as for flufenamic acid and flutamide. 45 theophylline, 50 flufenamic acid, 21 flutamide, and 117 cocaine candidates within 10 kJ/mol of the respective lowest-energy structure were retained and are considered in this work. They can be found (in CIF format) in the supplementary information of Ref. [13]. The **AZD8329** structures were geometry optimized using the molecular mechanics description outlined in Ref. [21], using the Open Force Field module of the Cerius2 v4.6 package, and refined using DMACRYS [26] with DFT calculations in the Gaussian03 software [27] for the intra-molecular contribution and an atom-atom model of inter-molecular interactions with atomic multipole electrostatics. 11 AZD8329 candidates within 30 kJ/mol of the most stable predicted crystal structure for a given conformation were further geometry optimised using CASTEP [28] at the PBE-DFT level of theory and can be found in the supplementary information of Ref. [14]. The **ampicillin** candidates were geometry optimised with fixed (gas phase) molecular geometry using DMACRYS with an atomic multipoles model for the inter-molecular interactions based on the B3LYP/6-311G** charge density. Candidates within 20 kJ/mol of the lowest energy structure (and retaining at least five candidates for each conformer irrespective of relative stability) were refined using geometry optimisations performed in the CASTEP suite with the PBE functional and a Grimme D2 dispersion correction. The top ampicillin 23 candidates in terms of the RMSD of their chemical shifts with respect to experiment are considered in this work and can be found in the supplementary information of Ref. [12].

B. GIPAW DFT calculations of NMR response

The GIPAW DFT calculations for the different compounds were performed as follows:

- **Flutamide and theophylline:** the NMR calculations were performed using CASTEP v5.0 with the PBE exchange-correlation functional [29] without dispersion correction, an equivalent plane-wave energy cut-off of 550 eV and a Monkhorst-Pack k-point grid [30] with a maximum spacing of $2\pi \times 0.05\text{Å}^{-1}$. The calculations used on-the-fly generated GIPAW pseudopotentials [5].

- **Flufenamic acid:** the NMR calculations were performed using CASTEP v5.5 with the PBE exchange-correlation functional [29] with a Tkatchenko-Scheffler semi-empirical dispersion correction [31], an equivalent plane-wave energy cut-off of 700 eV and a Monkhorst-Pack k-point grid with a maximum spacing of $2\pi \times 0.05\text{Å}^{-1}$. The calculations used on-the-fly generated GIPAW pseudopotentials.

- **Ampicillin, AZD8320, and cocaine:** the NMR calculations were performed using Quantum Espresso v6.3. with the PBE exchange-correlation functional [29] with a Grimme D2 semi-empirical dispersion correction [32, 33] and an equivalent plane-wave energy cut-off of 100 and 400 Rydberg for the wavefunction and density, respectively. The calculations used pseudopotentials from the PSLibrary [34]
  H.pbe-kjpaw_psl.1.0.0.UPF,
  C.pbe-n-kjpaw_psl.1.0.0.UPF,
  N.pbe-n-kjpaw_psl.1.0.0.UPF,
  O.pbe-nl-kjpaw_psl.1.0.0.UPF,
  S.pbe-nl-kjpaw_psl.1.0.0.UPF.

The full set of candidate structures and the associated GIPAW and ML shifts are available as supplementary data [35].

C. Partial and full assignments of experimental NMR shifts to particular nuclei

In the main text we assume the following partial assignments of the measured NMR shifts to nuclei in the respective compounds, which can (conservatively) be made on the basis of the 1D spectra alone (the numbering scheme is indicated in Fig. S1):

- **Flutamide**.

  ^{1}H: (i) the aliphatic protons (10, 11 and 12) are between 0 and 4 ppm, (ii) the protons in the aromatic and amide groups (3, 5, 6, 8) are between 6 and 11 ppm.

  ^{13}C: (i) the peaks between 10 and 40 ppm are carbon 10, 11 and 12, (ii) the peak at 176 ppm is the carbonyl (9).

- **Flufenamic acid**.

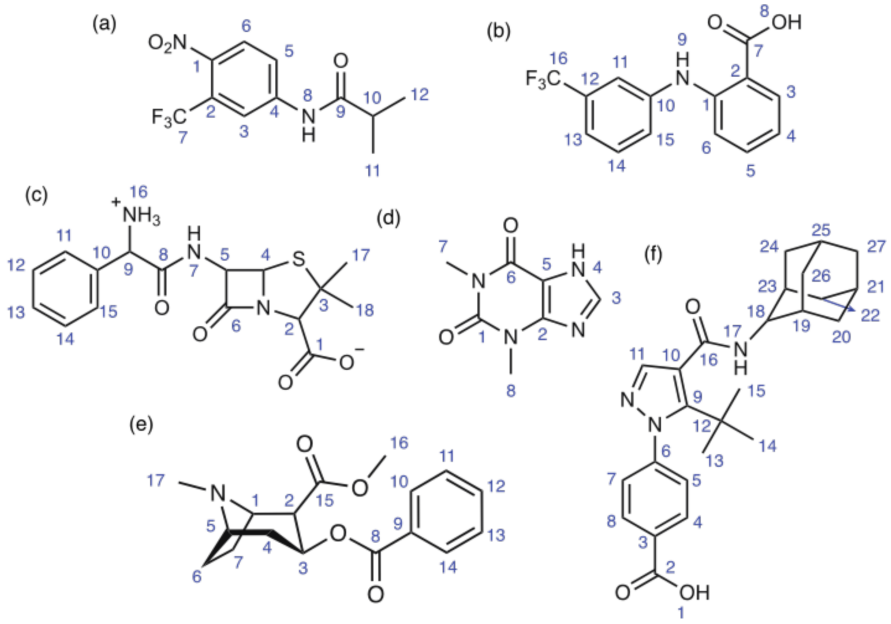

FIG. S1. Chemical structures of (a) flutamide, (b) flufenamic acid, (c) ampicillin, (d) theophylline, (e) cocaine, (f) AZD8329. The distinct ^{1}H sites are numbered.

^{1}H: Due to large overlap of the resonances, no assignments can safely be made on the basis of the 1D spectra alone.

^{13}C: (i) the peak at 175 ppm is the carbonyl (7).

- **Ampicillin**.

  ^{1}H: (i) the peaks between 0 and 2 ppm are the methyl groups (17, 18).

  ^{13}C: (i) the peaks between 25 and 35 ppm are the methyl groups (17,18), (ii) the peaks between 120 and 140 ppm are the aromatic carbons (10, 11, 12, 13, 14, 15), (iii) the peaks between 165 and 175 ppm are the carbonyls (1, 6, 8)

- **Theophylline**.

  ^{1}H: (i) the peaks between 2 and 5 ppm are the methyl groups (7, 8), (ii) the peak at 15 ppm is the NH (4).

  ^{13}C: (i) the peaks at 30 ppm are the methyl carbons (7,8).

- **Cocaine**.

  ^{1}H: Due to large overlap of the resonances, no assignments can safely be made on the basis of the 1D spectra alone.

  ^{13}C: (i) the peaks between 160 and 180 ppm are the two carbonyls (8 and 15), (ii) the peaks between 120 and 140 ppm are the aromatic carbon atoms (9,10,11,12,13,14).

- **AZD8329**.

  ^{1}H: Due to large overlap of the resonances, no assignments can safely be made on the basis of the 1D spectra alone.

  ^{13}C: (i) the peaks between 170 and 180 ppm are the two carbonyls (2 and 16), (ii) the peaks between 110 and 150 ppm are the aromatic carbon atoms (3, 4, 5, 6, 7, 8, 9, 10, 11), (iii) the peaks between 20 and 70 ppm are the aliphatic carbon atoms (12, 13, 14, 15, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27).

The fully assigned chemical shifts are reported in Table S1.

D. Effects of different strategies for translating chemical shieldings into shifts

As touched upon in the main text the choice of strategy for converting the originally predicted chemical shieldings in to shifts – using either predetermined regression parameters [36] or on-the-fly linear regressions – does affect the predicted chemical shifts for candidate structures and thereby the results of NMR crystal structure determinations. Any errors arising from the conversion reduce the resolving power of NMR crystallography, in particular if they are not accounted for in the estimate of the uncertainty of shift predictions with respect to experiment, $\{\sigma_j\}$.

Notably, Tab. S2 shows that for the compounds considered in this work the RMSD between the predicted shifts for the correct candidate structures and the experimentally measured NMR shifts, are consistently higher when using the reference parameters obtained in Ref. [36] to

(a)

Flutamide

site	^{1}H shift	^{13}C shift
1	–	140.9 or 145.4
2	–	124.4
3	7.1 or 8.1	116.7 or 130.9
4	–	145.4 or 140.9
5	9.9	124.4
6	8.1 or 7.1	130.9 or 116.7
7	–	122.0
8	8.0	–
9	–	176.1
10	2.0	35.9
11	1.2	17.8 or 21.8
12	1.2	21.8 or 17.8

(b)

Flufenamic acid

site	^{1}H shift	^{13}C shift
1	–	149.3
2	–	109.7
3	8.3	133.0
4	6.0, 6.9 or 6.2	117.2, 121.7 or 119.8
5	5.4	136.3
6	6.8	112.0
7	–	175.0
8	12.4	–
9	9.6	–
10	–	139.9
11	6.9, 6.2 or 6.0	121.7, 119.8 or 117.2
12	–	131.7
13	6.2, 6.0 or 6.9	119.8, 117.2 or 121.7
14	5.9	129.5
15	7.3	128.1
16	–	124.1

(c)

Ampicillin

site	^{1}H shift	^{13}C shift
1	–	173.2
2	4.0	75.3
3	–	64.8
4	5.2	64.8
5	6.6	56.5
6	–	175.0
7	7.5	–
8	–	169.8
9	4.8	57.4
10	–	135.4
11	7.1, 7.2, 7.3,7.6 or 5.4	129.0, 132.0, 129.9, 126.9, or 128.3
12	5.4, 7.1, 7.2, 7.3 or 7.6	128.3, 126.9, 129.0, 132.0, or 129.9
13	7.6, 5.4, 7.1, 7.2 or 7.3	129.9, 128.3,126.9, 129.0 or 132.0
14	7.3, 7.6, 5.4, 7.1 or 7.2	132.0, 129.9, 128.3,126.9 or 129.0
15	7.2 , 7.3, 7.6, 5.4 or 7.1	129.0, 132.0, 129.9, 128.3 or 126.9
16	10.0	–
17	0.6	30.1
18	1.6	28.9

(d)

Theophylline

site	^{1}H shift	^{13}C shift
1	–	150.8
2	–	146.1
3	7.7	140.8
4	14.6	–
5	–	105.8
6	–	155.0
7	3.4	29.9
8	3.4	29.9

(e)

Cocaine

site	^{1}H shift	^{13}C shift
1	3.76	65.95
2	3.78	50.16
3	5.63	66.17
4	3.06 or 3.32	36.66
5	3.49	62.63
6	3.38 or 2.91	25.62
7	2.25/2.12	25.62
8	–	165.94
9	–	129.37
10	8.01	131.50
11	8.01	133.50
12	8.01	134.53
13	8.01	133.50
14	8.01	131.50
15	–	172.18
16	3.78	50.16
17	1.04	41.52

(f)

AZD8329

site	^{1}H shift	^{13}C shift
1	15.37	–
2	–	171.04
3	–	131.19
4	8.69	130.48 or 128.05
5	6.92	128.05 or 130.48
6	–	147.31
7	8.47	128.05 or 130.48
8	9.01	130.48 or 128.05
9	–	148.71
10	–	114.10
11	7.73	138.43
12	–	33.42
13	0.73	29.53
14	0.73	29.53
15	0.73	29.53
16	–	172.98
17	9.64	–
18	2.90	60.16
19	1.54	34.14
20	0.44 or 1.6	30.80 or 37.41
21	1.00	27.81
22	0.80	36.42 or 30.80
23	1.78	32.45
24	1.88	30.80 or 36.42
25	–	27.81
26	1.88	37.41 or 30.80
27	1.74	37.41

TABLE S1. Experimental ^{1}H and ^{13}C chemical shifts for (a) flutamide as reported in Ref. [13], (b) flufenamic acid as reported in Ref. [13], (c) ampicillin as reported in Ref. [12], (d) theophylline as reported in Ref. [12], (e) cocaine as reported in Ref. [13], and (f) AZD8329 as reported in Ref. [14].

convert shieldings into shifts than when employing on-the-fly linear regressions. While we have no reason to

	Predetermined [36]		On-the-fly	
compound	^{1}H	^{13}C	^{1}H	^{13}C
ampicillin	0.36	2.58	0.33	2.17
AZD8329	0.39	2.34	0.36	1.21
cocaine	0.41	2.39	0.31	1.44
theophylline	0.27	1.92	0.13	1.93
flufenamic acid	1.33	5.01	0.17	5.01
flutamide	0.54	3.84	0.37	3.76

TABLE S2. RMSD between GIPAW DFT predicted ^{1}H and ^{13}C shifts for the correct candidate structures (determined on the basis of the originally predicted chemical shieldings using either predetermined regression parameters [36] or on-the-fly linear regression) and the experimentally measured NMR shifts.

believe that this should generalize to other compounds, we note that the reference parameters are system dependent and vary noticeably between compounds [36], motivating data-driven approaches such as on-the-fly linear regression. Notably, the latter effectively eliminates two degrees of freedom in NMR crystal structure determinations, which might be crucial for compounds with few distinct chemical shifts.

It is worth pointing out that in assessing the errors $\{\sigma_j\}$ in shift predictions in a data-driven fashion by maximizing $p(\mathbf{y}^\star)$ with respect to $\{\sigma_j\}$ (see main text, appendix A) any errors arising from the conversion are absorbed into the (global) errors $\{\sigma_j\}$. In consequence, the resultant (confidences in) structure determinations are less affected by the choice of strategy for converting chemical shieldings into shifts. One possible strategy to minimize the impact of on-the-fly conversion of shieldings into shifts might be devised by simultaneously optimizing the linear regression coefficients and the uncertainty estimates by likelihood maximisation, analogous to Section 2.4.

E. Full sets of RMSDs of shifts with respect to experiment, KPCA maps of candidate similarity as seen through the lense of NMR experiments, and PCA maps of the structural similarity of the candidates

Figs. S4 to S17 show the complete sets of RMSDs of shifts with respect to experiment, KPCA maps of candidate similarity as seen through the lense of NMR experiments, and PCA maps of the structural similarity of the candidates. The right hand panels of Figs. S4 and S5 show RMSD differences between the predicted ^{1}H and ^{13}C shifts of the different CSP candidates and experiment. To be able to evaluate RMSDs in which differences in ^{1}H shifts are not completely outweighed by differences in ^{13}C shifts due to their larger absolute values and thus errors, we normalise all shifts by dividing them by the typical errors of GIPAW predictions with respect to experiment of 0.33 ppm for ^{1}H and 1.9 ppm for ^{13}C shifts. In consequence RMSDs are consistent with experiment if they lie within around 50% of a value of one.

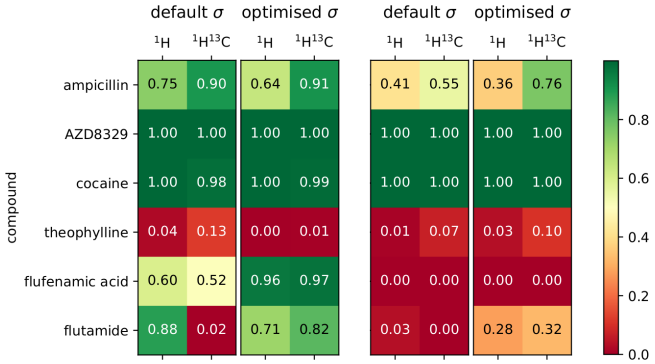

FIG. S2. Comparison of the Bayesian probabilities of matching experiment assigned to the correct CSP candidates when using on-the-fly system-specific linear regressions (left panel) and global reference parameters [36] (right panel) to convert the predicted chemical shieldings into shifts, respectively. In each case the probabilities are evaluated on the basis of the default global uncertainties (left-hand columns) and uncertainties estimated for each individual compound by maximizing $p(\mathbf{y}^\star)$ with respect to $\{\sigma_j\}$ as described in appendix A of the main text (right-hand columns).

Figs. S7, S9, S11, S13, S15, and S17 show two-dimensional representations of the structural similarity of the sets of CSP candidate structures as measured on the basis of their respective SOAP feature vectors. In each case the PCA is performed on the full set of CSP candidates, so that the two-dimensional projections of candidates (i.e. their positions on the similarity maps) do not change, irrespective of the availability of assignments of shifts to particular nuclei. Any differences in the figures arise solely from different sets of top 10 candidates depending on the availability of assignments of shifts to particular nuclei and resultant differences in the choice of origin and limits of the PCA coordinate axes.

FIG. S3. Overview of the results of NMR crystal structure determinations for ampicillin, AZD8329, cocaine, theophylline, flufenamic acid, and flutamide based on different degrees of experimental assignments of NMR shifts to nuclei and using shifts from ^{1}H and ^{13}C calculated with ML or DFT, respectively. Both full (fully assigned) and partial assignments (partially assigned) are detailed in Supplementary Section SII. Each cell is colored and labeled according to the Bayesian probability of matching experiment assigned to the representative of the experimental structure among the CSP candidates – this probability provides the key indicator of the reliability of the structure determination.

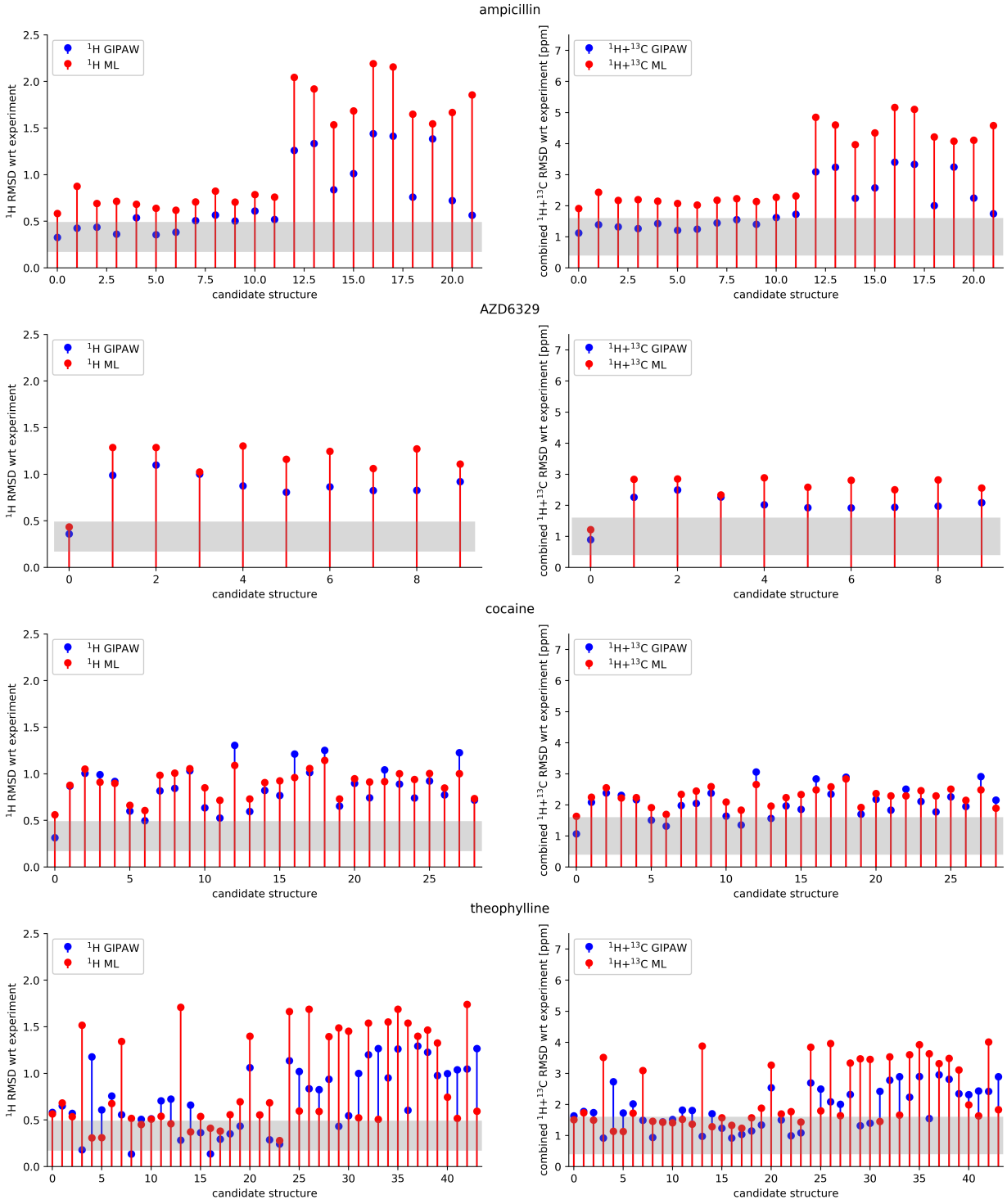

FIG. S4. RMSDs of the assigned GIPAW (blue) and ML (red) ^{1}H (left panel) and ^{1}H+^{13}C (right panel) shifts with respect to the experimentally measured shifts. The RMSDs for the structure determined by independent XRD measurements calculated using GIPAW and ML calculations are shown in cyan and magenta, respectively. The RMSDs for the combined ^{1}H and ^{13}C shifts are obtained by rescaling all shifts according to the estimates of the inherent errors in GIPAW predictions of 0.33 ± 0.16 ppm for ^{1}H and 1.9 ± 0.4 ppm for ^{13}C [36–38].

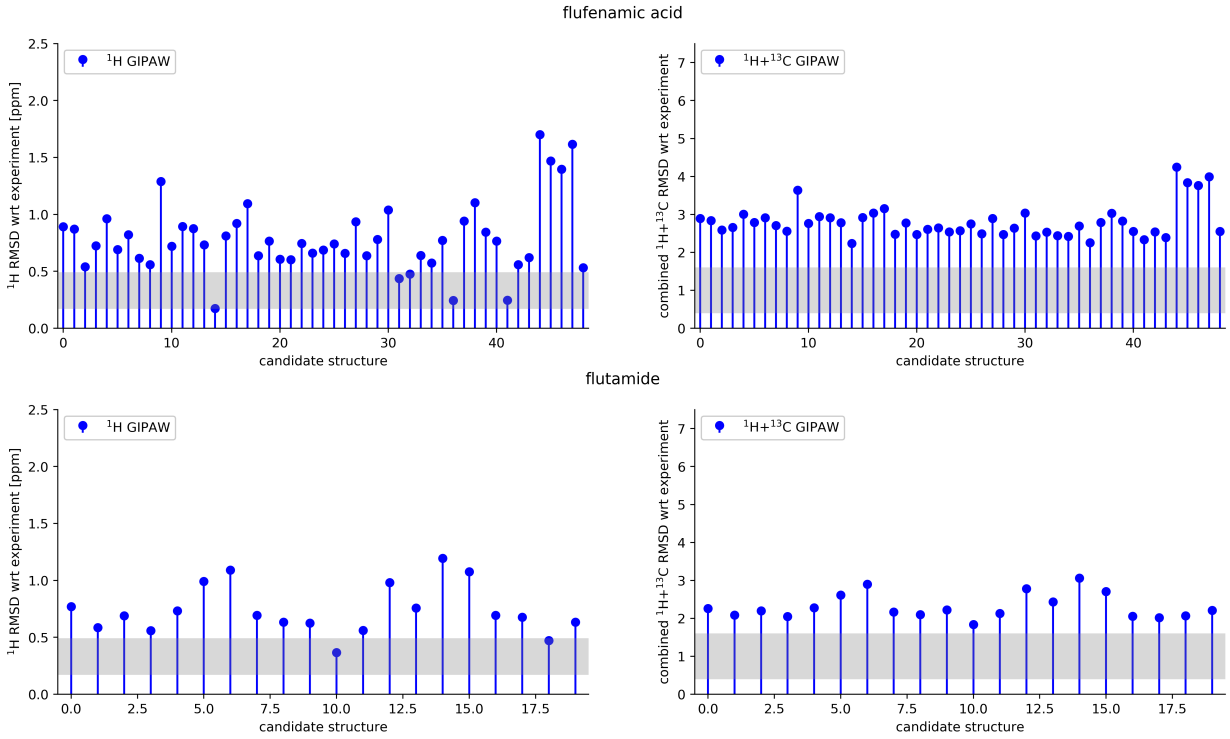

FIG. S5. RMSDs of the assigned GIPAW ^{1}H (left panel) and ^{1}H+^{13}C (right panel) shifts with respect to the experimentally measured shifts. The RMSDs for the structure determined by independent XRD measurements calculated using GIPAW calculations are shown in cyan. The RMSDs for the combined ^{1}H and ^{13}C shifts are obtained by rescaling all shifts according to the estimates of the inherent errors in GIPAW predictions of 0.33 ± 0.16 ppm for ^{1}H and 1.9 ± 0.4 ppm for ^{13}C [36–38].

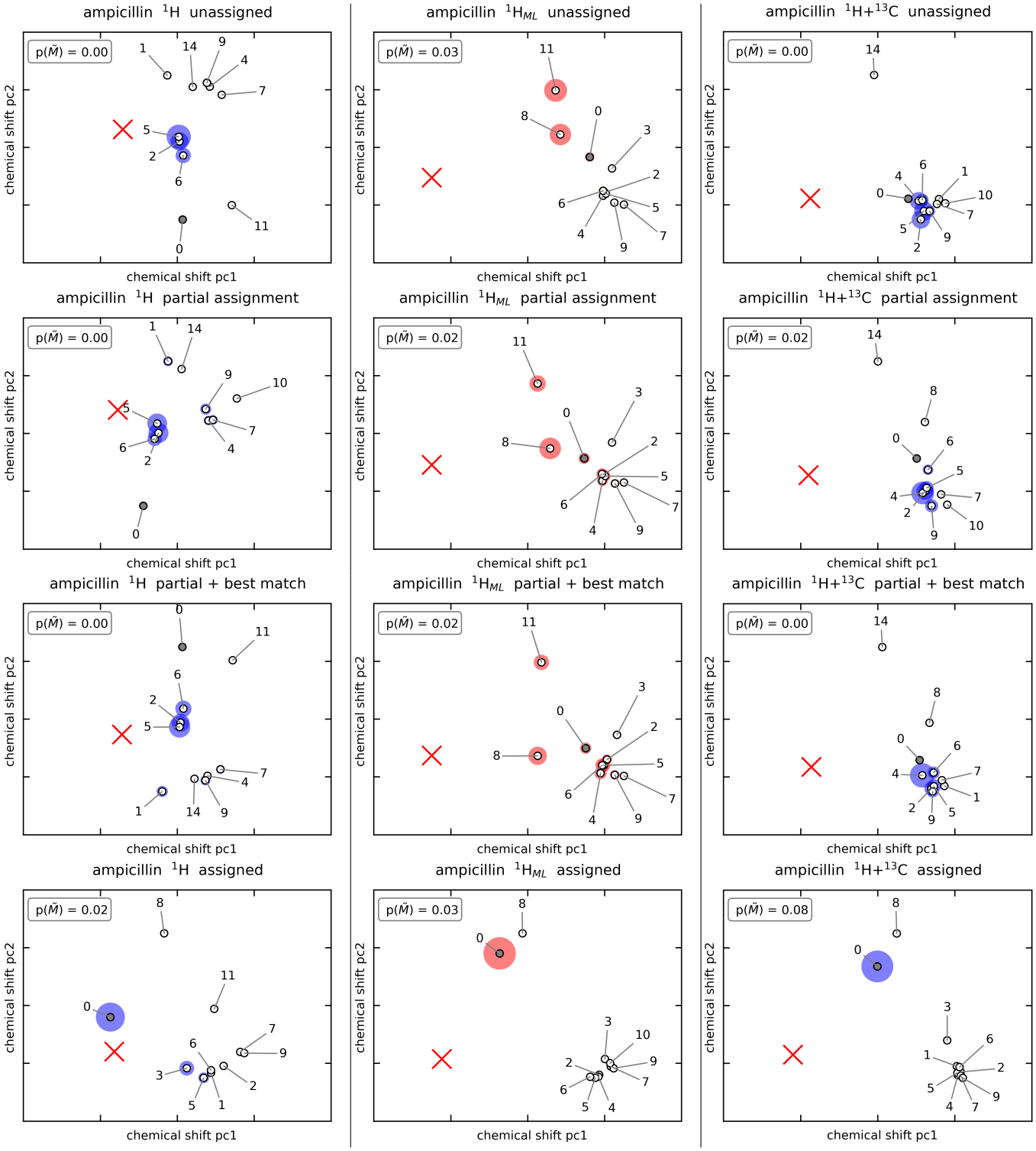

FIG. S6. Chemical shift similarity of the top 10 ampicillin candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts and (c) their SOAP structural features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the KPCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\left(M|\mathbf{y}^{\star}\right)$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

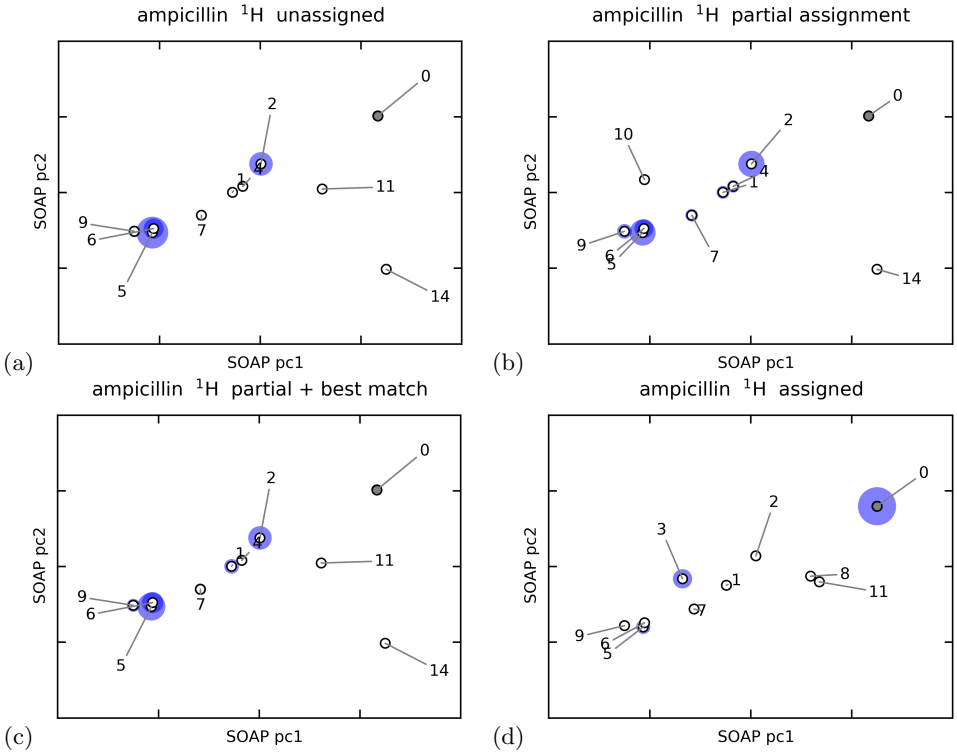

FIG. S7. Structural similarity of the top 10 ampicillin candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

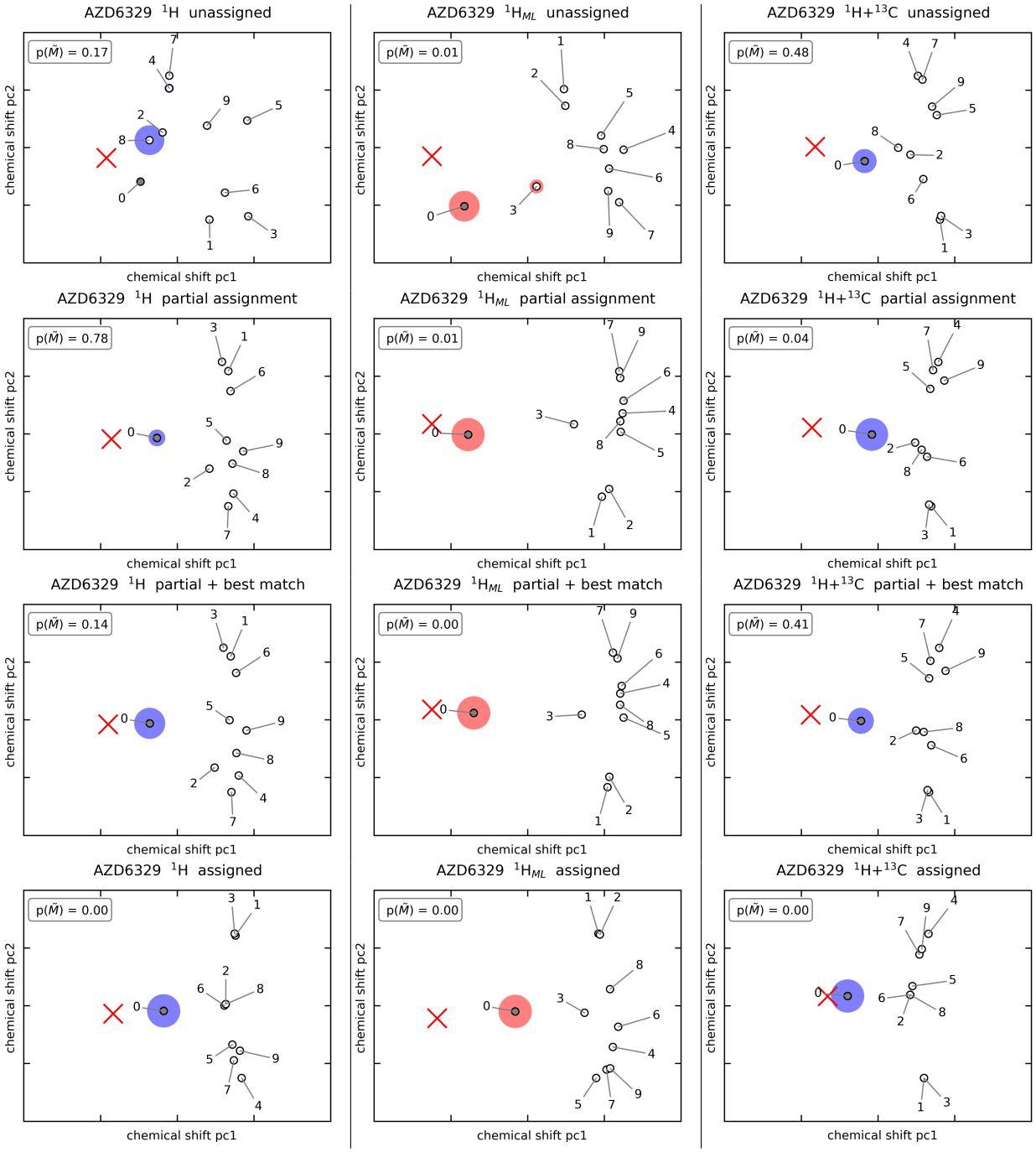

FIG. S8. Chemical shift similarity of the top 10 AZD8329 candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p(M|\mathbf{y}^{\star})$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

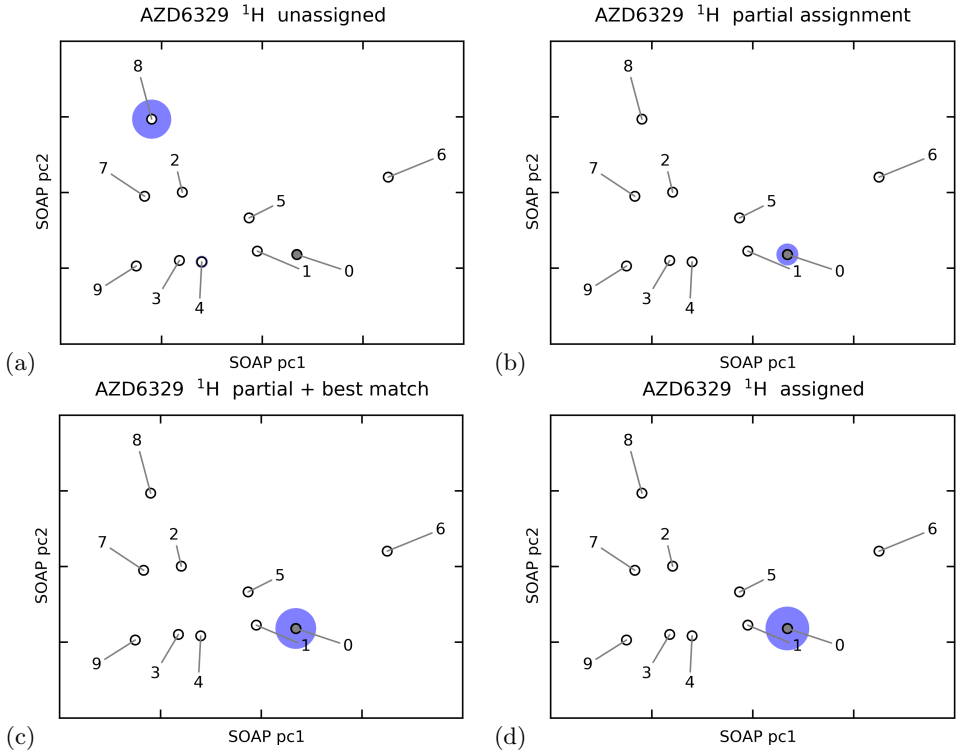

FIG. S9. Structural similarity of the top 10 azd candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

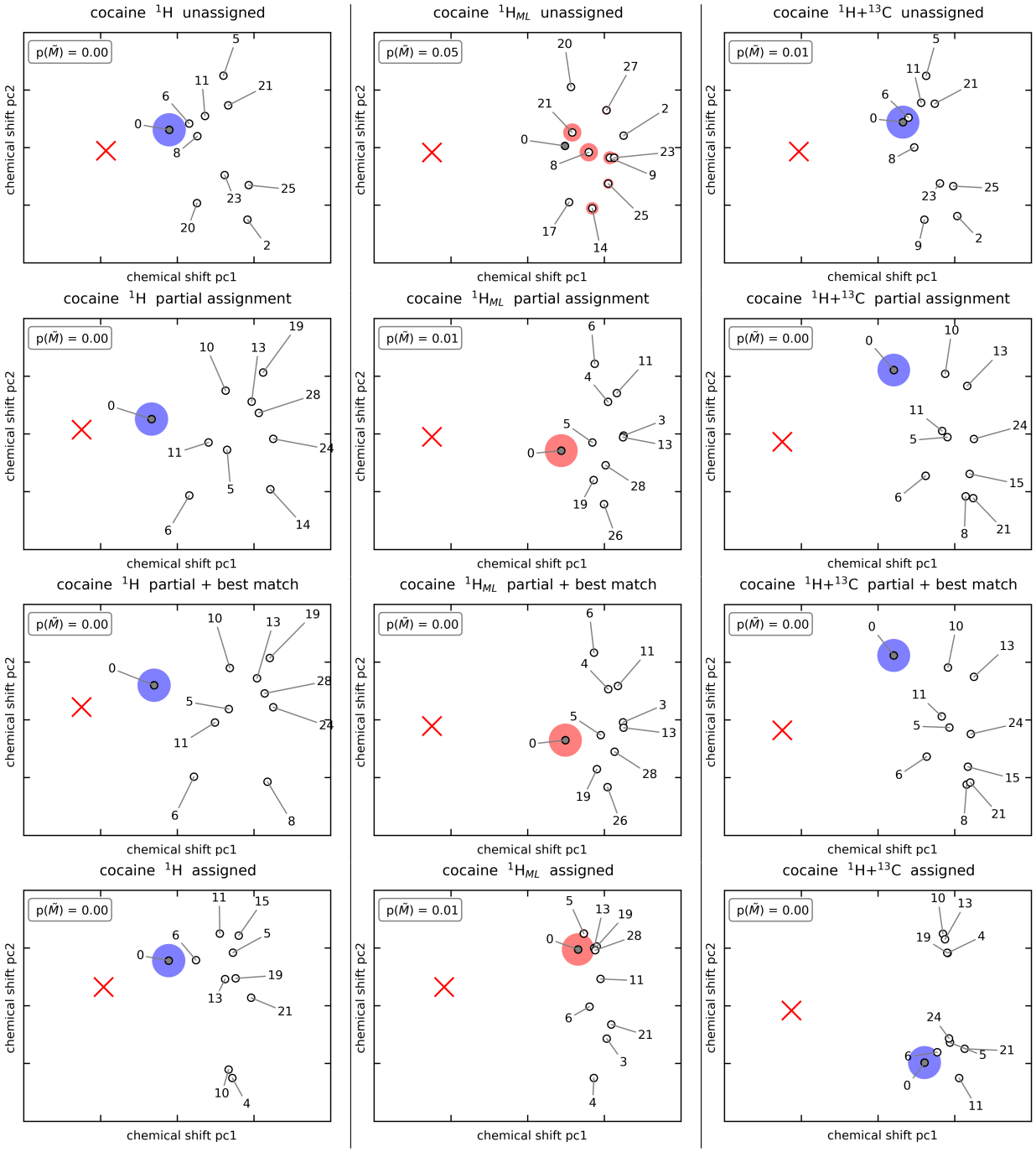

FIG. S10. Chemical shift similarity of the top 10 cocaine candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

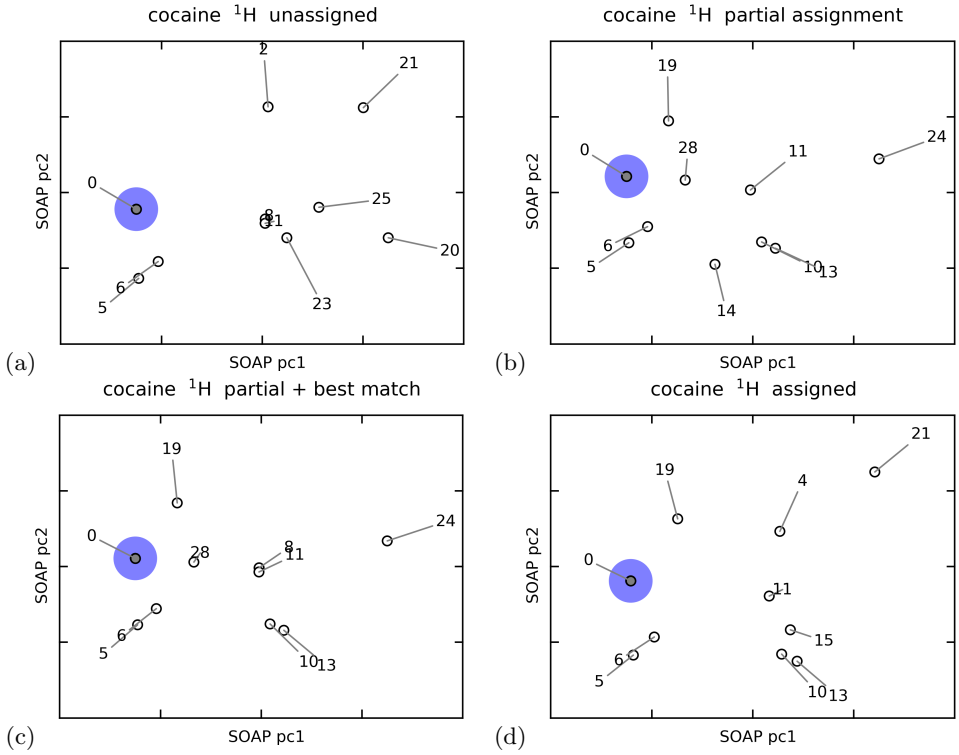

FIG. S11. Structural similarity of the top 10 cocaine candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

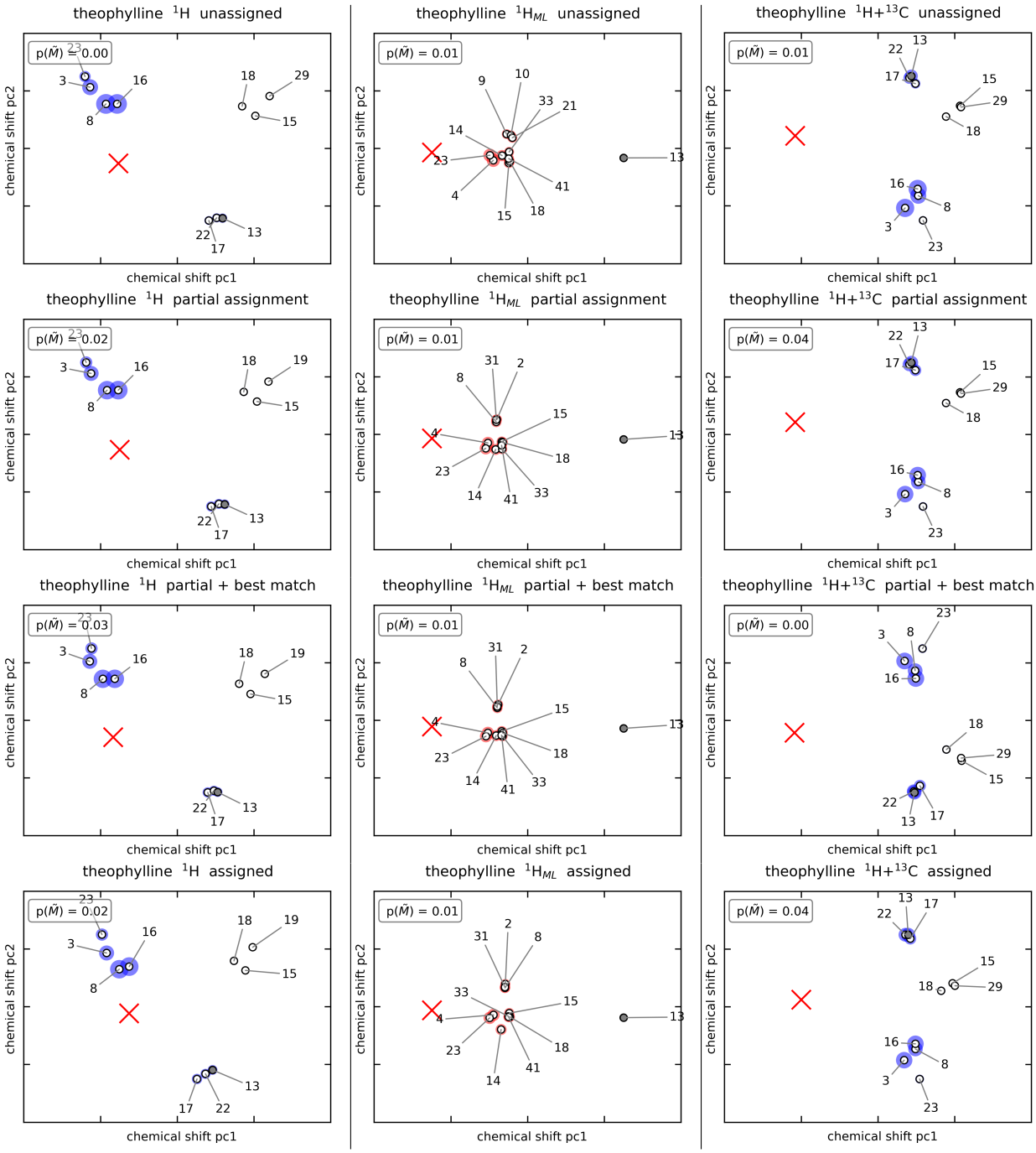

FIG. S12. Chemical shift similarity of the top 10 theophylline candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^{\star})$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

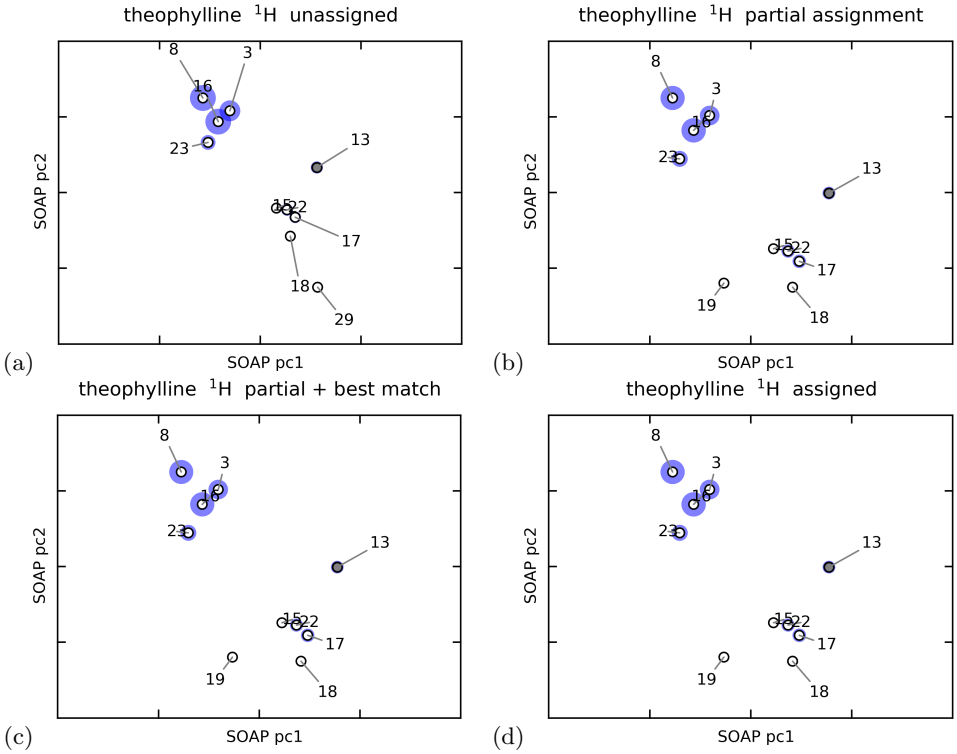

FIG. S13. Structural similarity of the top 10 theophylline candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

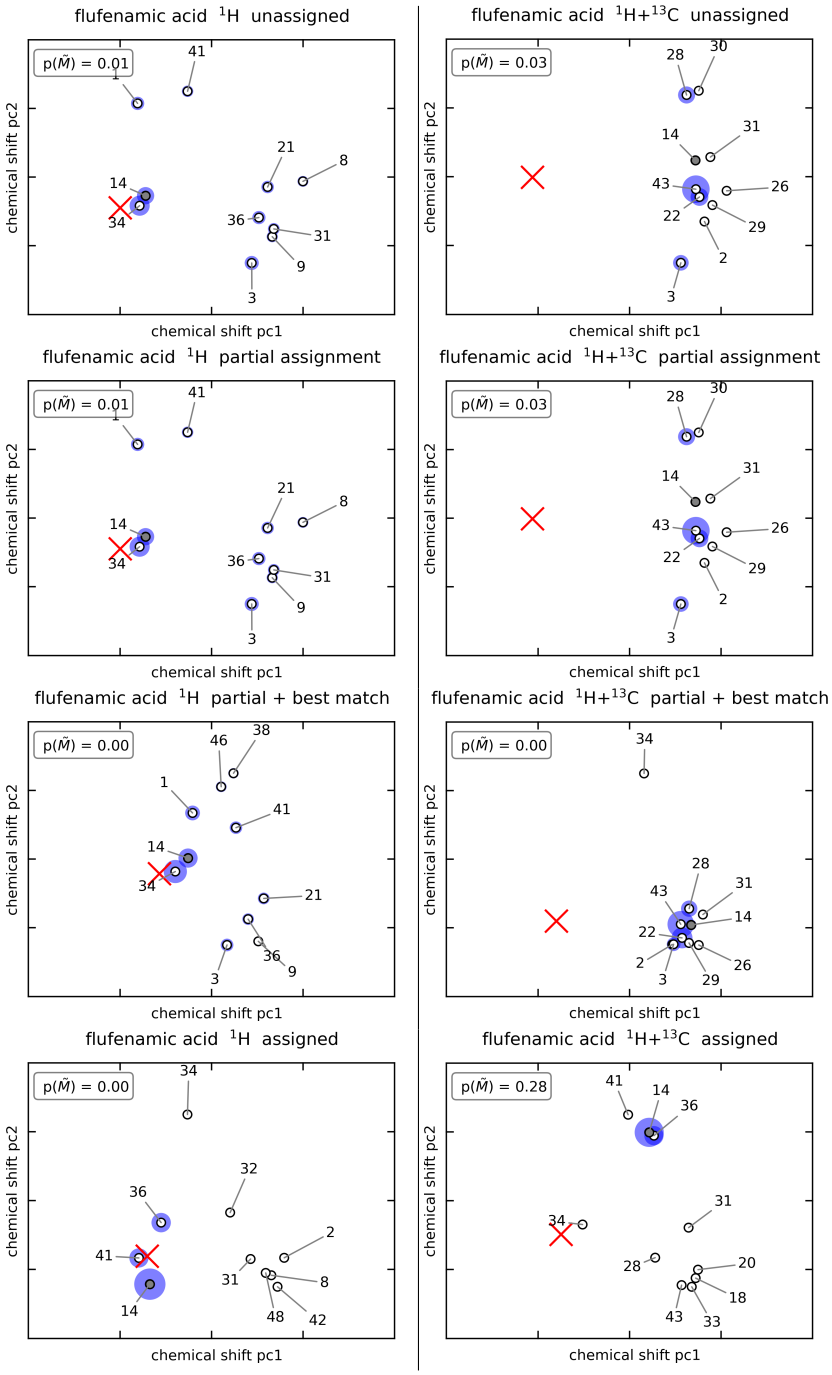

FIG. S14. Chemical shift similarity of the top 10 flufenamic acid candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^{\star})$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

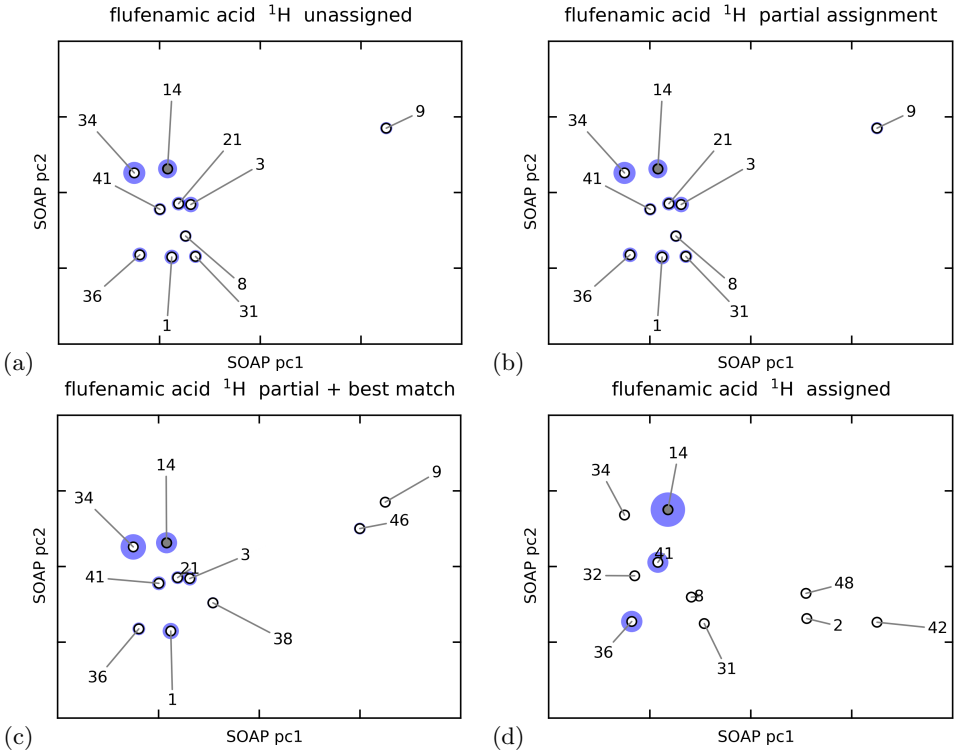

FIG. S15. Structural similarity of the top 10 flufenamic candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

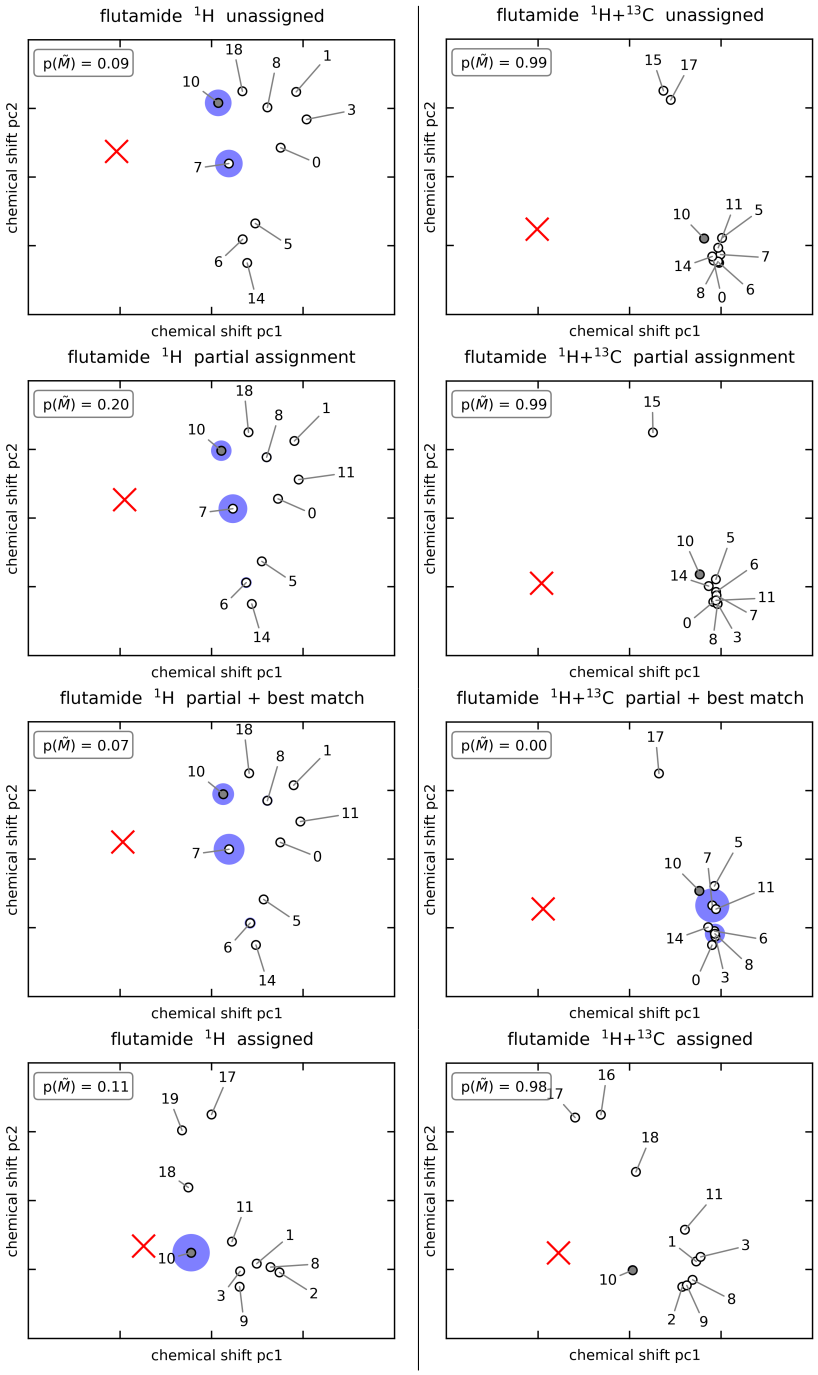

FIG. S16. Chemical shift similarity of the top 10 flutamide candidates in terms of (a) their unassigned and (b) assigned GIPAW ^{1}H shifts. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^{\star})$, are indicated by the area of the blue disks. Experiment is indicated by a red cross.

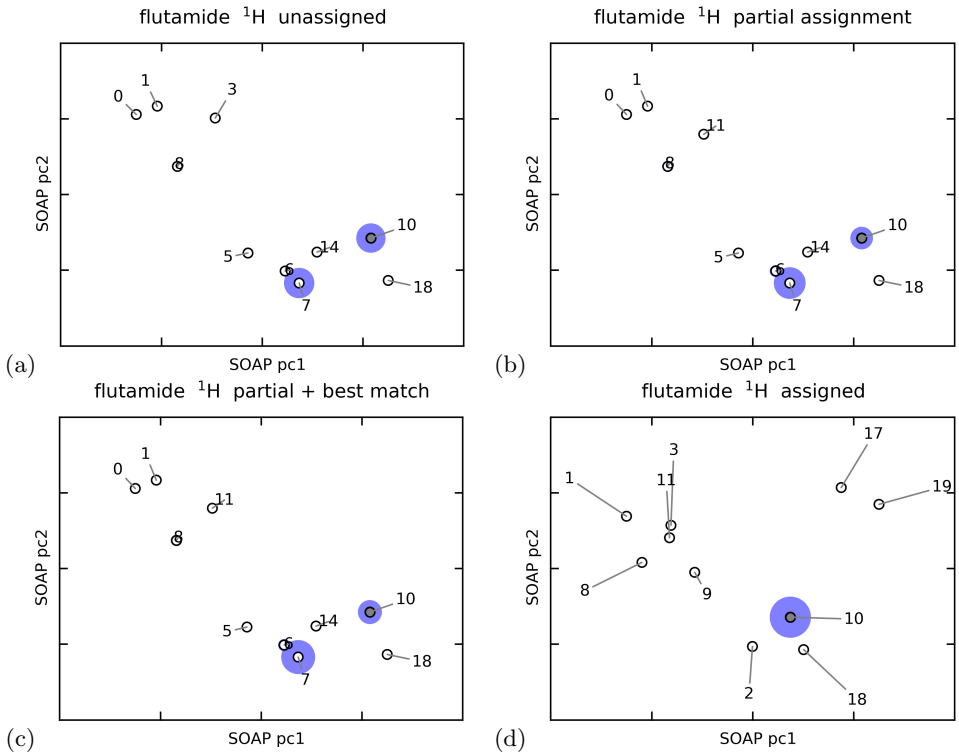

FIG. S17. Structural similarity of the top 10 flutamide candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts in terms of their SOAP features. The relative distance of structures are a measure of their (dis-)similarity. However, the absolute value of the abstract, collective principal components, pc1 and pc2, from the PCA constructions described in section III G has no intuitive physical meaning and is therefore not shown. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^\star)$, are indicated by the area of the blue disks.

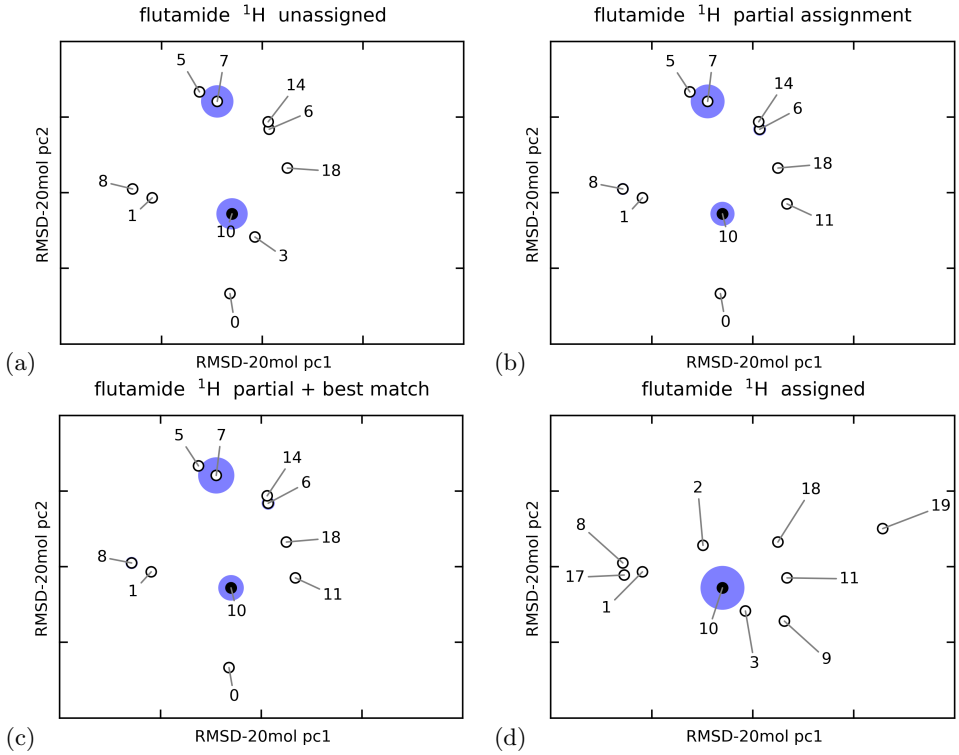

FIG. S18. Structural (dis-)similarity of the top 10 flutamide candidates according to (a) unassigned, (b) partially assigned, (c) best-match assigned, and (d) fully assigned GIPAW ^{1}H shifts as measured using the RMSD-1mol [39] metric of structural differences between candidates. The abstract, collective principal components, RMSD-1mol pc1 and pc2, are determined by performing a KPCA on the kernel matrix obtained by centering the matrix of RMSD-1mol distances between pairs of candidates. The candidates are labelled according to their rank in terms of configurational energy with the zero indicating the energetically most favourable candidate. The "true" and "false" CSP candidates are shown as filled in and empty circles, respectively. Their probabilities of matching experiment, $p\,(M|\mathbf{y}^{\star})$, are indicated by the area of the blue disks.

[1] H. W. Kuhn, Naval Research Logistics Quarterly **2**, 83 (1955).
[2] B. C. Gerstein, R. G. Pembleton, R. C. Wilson, and L. M. Ryan, Journal of Chemical Physics **66**, 361 (1977).
[3] Y. I. A. R. Nishiyama, M. Malon, Journal of Magnetic Resonance **244**, 1 (2014).
[4] V. Agarwal, S. Penzel, K. Szekely, R. Cadalbert, E. Testori, A. Oss, J. Past, A. Samoson, M. Ernst, A. Böckmann, and B. H. Meier, Angewandte Chemie International Edition **53**, 12253 (2014).
[5] C. J. Pickard and F. Mauri, Physical Review B **63**, 245101 (2001).
[6] J. R. Yates, C. J. Pickard, and F. Mauri, Physical Review B **76**, 024401 (2007).
[7] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, Nature Communications **9**, 4501 (2018).
[8] A. Lesage, C. Auger, S. Caldarelli, and L. Emsley, Journal of the American Chemical Society **119**, 7867 (1997).
[9] A. J. Rossini, A. Zagdoun, F. Hegner, M. Schwarzwalder, D. Gajan, C. Copéret, A. Lesage, and L. Emsley, Journal of the American Chemical Society **134**, 16899 (2012).
[10] A. Lesage, D. Sakellariou, S. Steuernagel, and L. Emsley, Journal of the American Chemical Society **120**, 13194 (1998).
[11] A. Lesage and L. Emsley, Journal of Magnetic Resonance **148**, 449 (2001).
[12] A. Hofstetter, M. Balodis, F. M. Paruzzo, C. M. Widdifield, G. Stevanato, A. Pinon, P. J. Bygrave, G. M. Day, and L. Emsley, Journal of the American Chemical Society **t**, t (2019).
[13] M. Baias *et al.*, Phys. Chem. Chem. Phys. **15**, 8069 (2013).
[14] M. Baias, J. N. Dumez, P. H. Svensson, S. Schantz, G. M. Day, and L. Emsley, Journal of the American Chemical Society **135**, 17501 (2013).
[15] A. C. Pinon, A. J. Rossini, C. M. Widdifield, D. Gajan, and L. Emsley, Molecular Pharmaceutics **12**, 4146 (2015).
[16] P. G. Karamertzanis and C. C. Pantelides, Journal of Computational Chemistry **26**, 304 (2005).
[17] D. H. Case, J. E. Campbell, P. J. Bygrave, and G. M. Day, Journal of Chemical Theory and Computation **12**, 910 (2016).
[18] A. D. Becke, Journal of Chemical Physics **98**, 5648 (1993).
[19] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, Journal of Physical Chemistry **98**, 11623 (1994).
[20] I. Kolossváry and W. C. Guida, Journal of the American Chemical Society **118**, 5011 (1996).
[21] G. M. Day, W. D. S. Motherwell, and W. Jones, Physical Chemistry Chemical Physics **9**, 1693 (2007).
[22] S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, and G. M. Day, Physical Chemistry Chemical Physics **12**, 8478 (2010).
[23] D. S. Coombes, S. L. Price, D. J. Willock, and M. Leslie, Journal of Physical Chemistry **100**, 7352 (1996).
[24] A. J. Stone and M. Alderton, Molecular Physics **56**, 1047 (1985).
[25] A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, and C. C. Pantelides, Journal of Chemical Theory and Computation **7**, 1998 (2011).
[26] S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, and G. M. Day, Physical Chemistry Chemical Physics **12**, 8478 (2010).
[27] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, "Gaussian 03, Revision C.02," Gaussian, Inc., Wallingford, CT, 2004.
[28] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne, Zeitschrift für Kristallographie **220**, 567 (2005).
[29] J. P. P. K. B. M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).
[30] H. J. Monkhorst and J. D. Pack, Physical Review B **13**, 5188 (1976).
[31] A. Tkatchenko and M. Scheffler, Physical Review Letters **102**, 073005 (2009).
[32] S. Grimme, Journal of Computational Chemistry **27**, 1787 (2006).
[33] V. Barone *et al.*, Journal of Computational Chemistry **30**, 934 (2009).
[34] A. D. Corso, Computational Material Science **95**, 337 (2014).
[35] Supplementary information is available under...
[36] J. D. Hartman, R. A. Kudla, G. M. Day, L. J. Mueller, and G. J. O. Beran, Phys. Chem. Chem. Phys. **18**, 21686 (2016).
[37] E. Salager, G. M. Day, R. S. Stein, C. J. Pickard, B. Elena, and L. Emsley, J. Am. Chem. Soc. **132**, 2564 (2010).
[38] M. Dracinsky, P. Unzueta, and G. J. O. Beran, Physical Chemistry Chemical Physics **21**, 14992 (2019).
[39] J. A. Chisholm and S. Motherwell, J. Appl. Crystallogr. **38**, 228 (2005).