Supporting Information

Chemical Shifts to Metabolic Pathways: Identifying Metabolic Pathways Directly from a Single 2D NMR Spectrum

Abhinav Dubey[1, 2], Annapoorni Rangarajan[3], Debnath Pal[1, 4] *, Hanudatta S. Atreya[2, 5] *

[1] IISc Mathematics Initiative, Indian Institute of Science, Bangalore 560012, India

[2] NMR Research Centre, Indian Institute of Science, Bangalore 560012, India

[3] Molecular Reproduction, Development and Genetics, Indian Institute of Science, Bangalore 560012, India

[4] Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560012, India

[5] Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore 560012, India

* Corresponding authors: Debnath Pal [dpal@serc.iisc.ernet.in]

Hanudatta S. Atreya [hsatreya@sif.iisc.ernet.in]

Table of Contents

Text S1: Derivation of index score for pathways

We use the example depicted in Figure 2 (main text). The index scores shown in Figure 2(c) are calculated as follows:

Let I(*p*) denote index of peak *p*. In the example shown, *p* belongs to the set: $\{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$

The following parameters are used for indexing the grid:

1. Lower limit on ^{1}H chemical shift: $\boldsymbol{h_l}$,
2. Upper limit on ^{1}H chemical shift: $\boldsymbol{h_u}$,
3. Bin size in ^{1}H dimension: $\boldsymbol{\Delta h}$,
4. Lower limit on ^{13}C chemical shift: $\boldsymbol{c_l}$,
5. Upper limit on ^{13}C chemical shift: $\boldsymbol{c_u}$,
6. Bin size in ^{13}C dimension: $\boldsymbol{\Delta c}$

In the example shown in Figure 2 (main text), we get, $I(s_1) = 2$, $I(s_2) = 20$, $I(s_3) = 9$, $I(s_4) = I(s_5) = 17$, $I(s_6) = 27$, $I(s_7)$ = NA (as it is outside the grid limits defined by above paramters $\boldsymbol{h_l}$, $\boldsymbol{h_u}$, $\boldsymbol{c_l}$, $\boldsymbol{c_u}$)

Let $P_X(I(p))$ denote the index score of peak '*p*' for pathway X. In this case X is from set of pathways:{A, B, C}. This is computed as reciprocal of number of pathways to which I(*p*) belongs. Thus,

$P_A(2) = 1/2$, $P_A(9) = 1$, $P_A(17) = 1/2$, $P_A(20) = 1/3$, $P_A(27) = 0$

$P_B(2) = 0$, $P_B(9) = 0$, $P_B(17) = 1/2$, $P_B(20) = 1/3$, $P_B(27) = 1/2$

$P_C(2) = 1/2$, $P_C(9) = 0$, $P_C(17) = 0$, $P_C(20) = 1/3$, $P_C(27) = 1/2$

The score for each pathway is normalized $P'_X(I(p))$ to using following equation

$$P'_X(I(p)) = \frac{P_X(I(p))}{\sum_{i \epsilon I(p)} P_X(i)}$$

We get,

P'$_A$(2) = 3/14, P'$_A$(9) = 3/7, P'$_A$(17) = 3/14, P'$_A$(20) = 1/7, P'$_A$(27) = 0 as $\sum_{i \epsilon I(p)} P_A(i) = \frac{7}{3}$

P'$_B$(2) = 0, P'$_B$(9) = 0, P'$_B$(17) = 3/8, P'$_B$(20) = 1/4, P'$_B$(27) = 3/8 as $\sum_{i \epsilon I(p)} P_B(i) = \frac{4}{3}$

P'$_C$(2) = 3/8, P'$_C$(9) = 0, P'$_C$(17) = 0, P'$_C$(20) = 1/4, P'$_C$(27) = 3/8 as $\sum_{i \epsilon I(p)} P_C(i) = \frac{4}{3}$

Text S2: Computing p-value for statistical significance of pathways

Suppose we want to calculate statistical significance of presence of pathway ‘P’

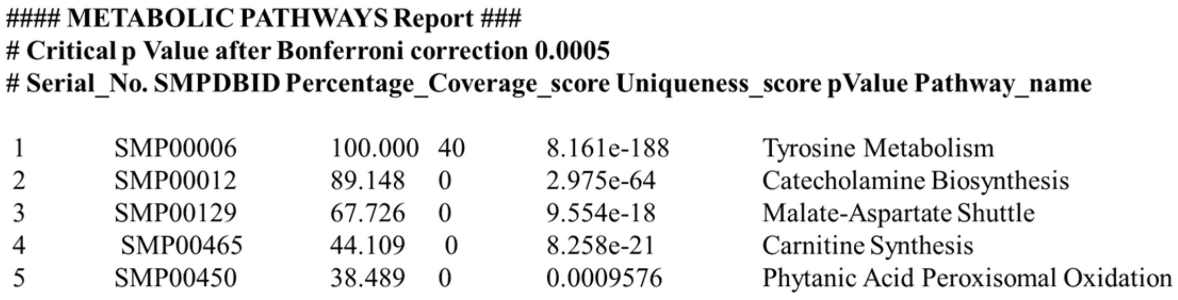
Let *N* be the total number of indices computed using entire chemical shifts database. Let x_p denotes the number of indices (out of total *N* indices) pointing to pathway ‘P’.

Let *M* be the total number of indices computed using experimental chemical shifts of sample. Let y_p denotes the number of indices (out of total *M* indices) pointing to pathway ‘P’.

Using hypergeometric distribution we calculated the propability to observe y_p/M by random chance given the proportion x_p/N of pathway ‘P’ in the database [1]. We also do ‘Bonferroni correction’ [2]to take into account observation of significant p-value due to multiple hypothesis testing.

We tested our method by giving peaklist of Tyrosine metabolism as input and computing the PC, PU and p-values.

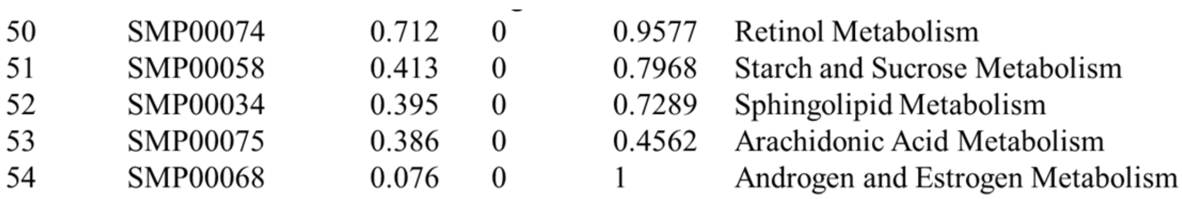
The top 5 pathways in the output are shown below



```
#### METABOLIC PATHWAYS Report ###
# Critical p Value after Bonferroni correction 0.0005
# Serial_No. SMPDBID Percentage_Coverage_score Uniqueness_score pValue Pathway_name

 1        SMP00006        100.000  40       8.161e-188       Tyrosine Metabolism
 2        SMP00012        89.148    0       2.975e-64        Catecholamine Biosynthesis
 3        SMP00129        67.726    0       9.554e-18        Malate-Aspartate Shuttle
 4         SMP00465       44.109     0      8.258e-21        Carnitine Synthesis
 5        SMP00450        38.489    0       0.0009576        Phytanic Acid Peroxisomal Oxidation
```

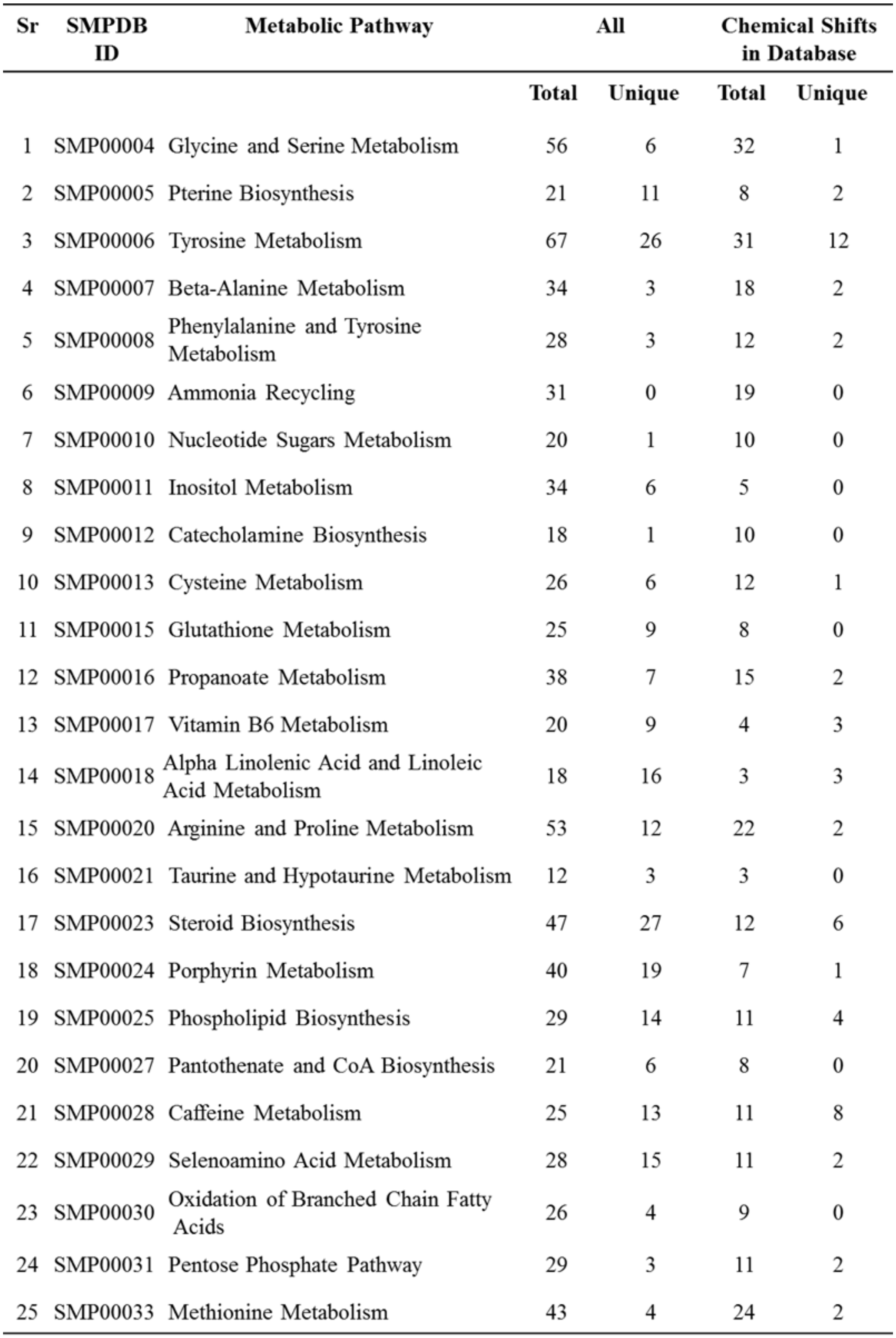
The bottom 5 pathways in the output are shown below



```
50    SMP00074    0.712   0    0.9577   Retinol Metabolism
51    SMP00058    0.413   0    0.7968   Starch and Sucrose Metabolism
52    SMP00034    0.395   0    0.7289   Sphingolipid Metabolism
53    SMP00075    0.386   0    0.4562   Arachidonic Acid Metabolism
54    SMP00068    0.076   0    1        Androgen and Estrogen Metabolism
```

(1) Rivals, I.; Personnaz, L.; Taing, L.; Potier, M. C. *Bioinformatics* **2007**, *23*, 401-407.
(2) Armstrong, R. A. *Ophthal Physl Opt* **2014**, *34*, 502-508.

Table S1: List of 91 SMPDB pathways used in ChemSMP along with their SMPDB IDs. The number of metabolites involved in each pathway and those unique to that pathway are shown in fourth and fifth column. Sixth and seventh column are subset of fourth and fifth column satisfying the condition that their 2D [^{13}C, ^{1}H] chemical shifts are available in HMDB database.



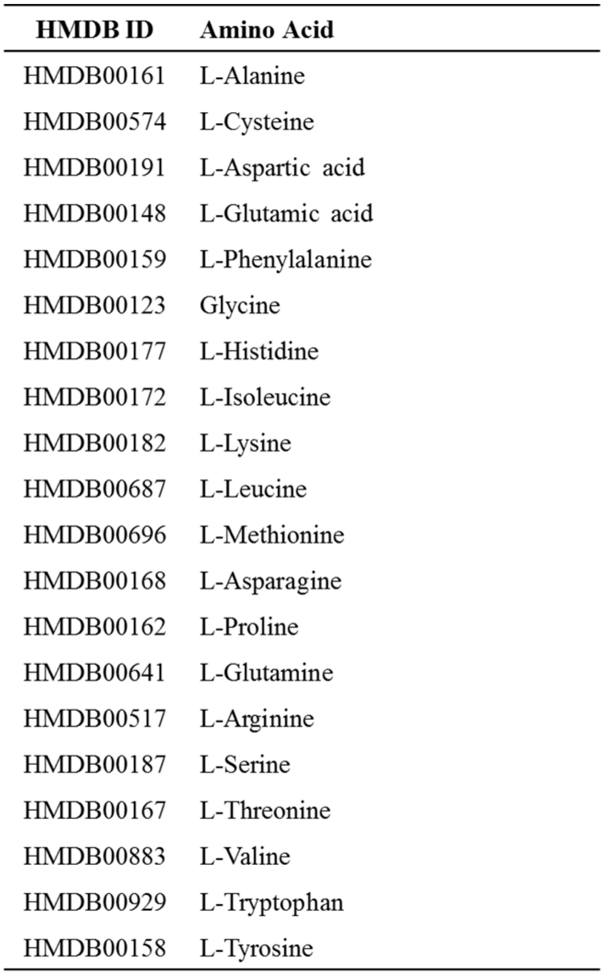
Sr	SMPDB ID	Metabolic Pathway	All		Chemical Shifts in Database	
			Total	Unique	Total	Unique
1	SMP00004	Glycine and Serine Metabolism	56	6	32	1
2	SMP00005	Pterine Biosynthesis	21	11	8	2
3	SMP00006	Tyrosine Metabolism	67	26	31	12
4	SMP00007	Beta-Alanine Metabolism	34	3	18	2
5	SMP00008	Phenylalanine and Tyrosine Metabolism	28	3	12	2
6	SMP00009	Ammonia Recycling	31	0	19	0
7	SMP00010	Nucleotide Sugars Metabolism	20	1	10	0
8	SMP00011	Inositol Metabolism	34	6	5	0
9	SMP00012	Catecholamine Biosynthesis	18	1	10	0
10	SMP00013	Cysteine Metabolism	26	6	12	1
11	SMP00015	Glutathione Metabolism	25	9	8	0
12	SMP00016	Propanoate Metabolism	38	7	15	2
13	SMP00017	Vitamin B6 Metabolism	20	9	4	3
14	SMP00018	Alpha Linolenic Acid and Linoleic Acid Metabolism	18	16	3	3
15	SMP00020	Arginine and Proline Metabolism	53	12	22	2
16	SMP00021	Taurine and Hypotaurine Metabolism	12	3	3	0
17	SMP00023	Steroid Biosynthesis	47	27	12	6
18	SMP00024	Porphyrin Metabolism	40	19	7	1
19	SMP00025	Phospholipid Biosynthesis	29	14	11	4
20	SMP00027	Pantothenate and CoA Biosynthesis	21	6	8	0
21	SMP00028	Caffeine Metabolism	25	13	11	8
22	SMP00029	Selenoamino Acid Metabolism	28	15	11	2
23	SMP00030	Oxidation of Branched Chain Fatty Acids	26	4	9	0
24	SMP00031	Pentose Phosphate Pathway	29	3	11	2
25	SMP00033	Methionine Metabolism	43	4	24	2

Sr	SMPDB ID	Metabolic Pathway	All		Chemical Shifts in Database	
			Total	Unique	Total	Unique
26	SMP00034	Sphingolipid Metabolism	40	18	14	3
27	SMP00035	Bile Acid Biosynthesis	64	45	15	5
28	SMP00036	D-Arginine and D-Ornithine Metabolism	11	4	1	1
29	SMP00037	Lysine Degradation	27	6	13	3
30	SMP00039	Glycerolipid Metabolism	25	2	16	1
31	SMP00040	Glycolysis	25	0	14	0
32	SMP00041	Sulfate/Sulfite Metabolism	23	8	6	2
33	SMP00043	Galactose Metabolism	36	8	22	4
34	SMP00044	Histidine Metabolism	40	8	19	4
35	SMP00045	Amino Sugar Metabolism	32	10	18	5
36	SMP00046	Pyrimidine Metabolism	59	21	28	14
37	SMP00048	Nicotinate and Nicotinamide Metabolism	38	10	14	4
38	SMP00050	Purine Metabolism	74	32	35	16
39	SMP00051	Fatty acid Metabolism	43	1	8	1
40	SMP00052	Beta Oxidation of Very Long Chain Fatty Acids	14	4	8	2
41	SMP00053	Folate Metabolism	29	6	9	0
42	SMP00054	Fatty Acid Elongation In Mitochondria	34	0	5	0
43	SMP00055	Alanine Metabolism	18	1	12	1
44	SMP00057	Citric Acid Cycle	34	2	19	1
45	SMP00058	Starch and Sucrose Metabolism	30	8	17	3
46	SMP00059	Urea Cycle	28	0	17	0
47	SMP00060	Pyruvate Metabolism	47	7	20	1
48	SMP00063	Tryptophan Metabolism	61	33	25	13
49	SMP00064	Fructose and Mannose Degradation	32	10	15	2
50	SMP00065	Ubiquinone Biosynthesis	19	10	3	1

Sr	SMPDB ID	Metabolic Pathway	All		Chemical Shifts in Database	
			Total	Unique	Total	Unique
51	SMP00066	Biotin Metabolism	7	2	4	1
52	SMP00067	Aspartate Metabolism	34	2	20	1
53	SMP00068	Androgen and Estrogen Metabolism	32	13	13	4
54	SMP00070	Riboflavin Metabolism	19	4	6	1
55	SMP00071	Ketone Body Metabolism	13	2	7	2
56	SMP00072	Glutamate Metabolism	50	3	25	2
57	SMP00073	Butyrate Metabolism	19	0	8	0
58	SMP00074	Retinol Metabolism	33	22	9	5
59	SMP00075	Arachidonic Acid Metabolism	71	58	5	2
60	SMP00076	Thiamine Metabolism	9	3	5	1
61	SMP00123	Betaine Metabolism	23	3	11	0
62	SMP00124	Glycerol Phosphate Shuttle	10	0	5	0
63	SMP00126	Phenylacetate Metabolism	9	3	5	1
64	SMP00127	Glucose-Alanine Cycle	11	0	8	0
65	SMP00128	Gluconeogenesis	35	0	20	0
66	SMP00129	Malate-Aspartate Shuttle	10	0	7	0
67	SMP00130	Steroidogenesis	43	32	11	5
68	SMP00355	Mitochondrial Electron Transport Chain	19	1	7	0
69	SMP00444	Lactose Synthesis	20	0	9	0
70	SMP00445	Spermidine and Spermine Biosynthesis	18	2	8	1
71	SMP00449	Ethanol Degradation	18	1	8	1
72	SMP00450	Phytanic Acid Peroxisomal Oxidation	24	1	9	0
73	SMP00452	Threonine and 2-Oxobutanoate Degradation	19	1	10	1
74	SMP00455	Homocysteine Degradation	8	0	6	0
75	SMP00456	Fatty Acid Biosynthesis	35	24	14	6

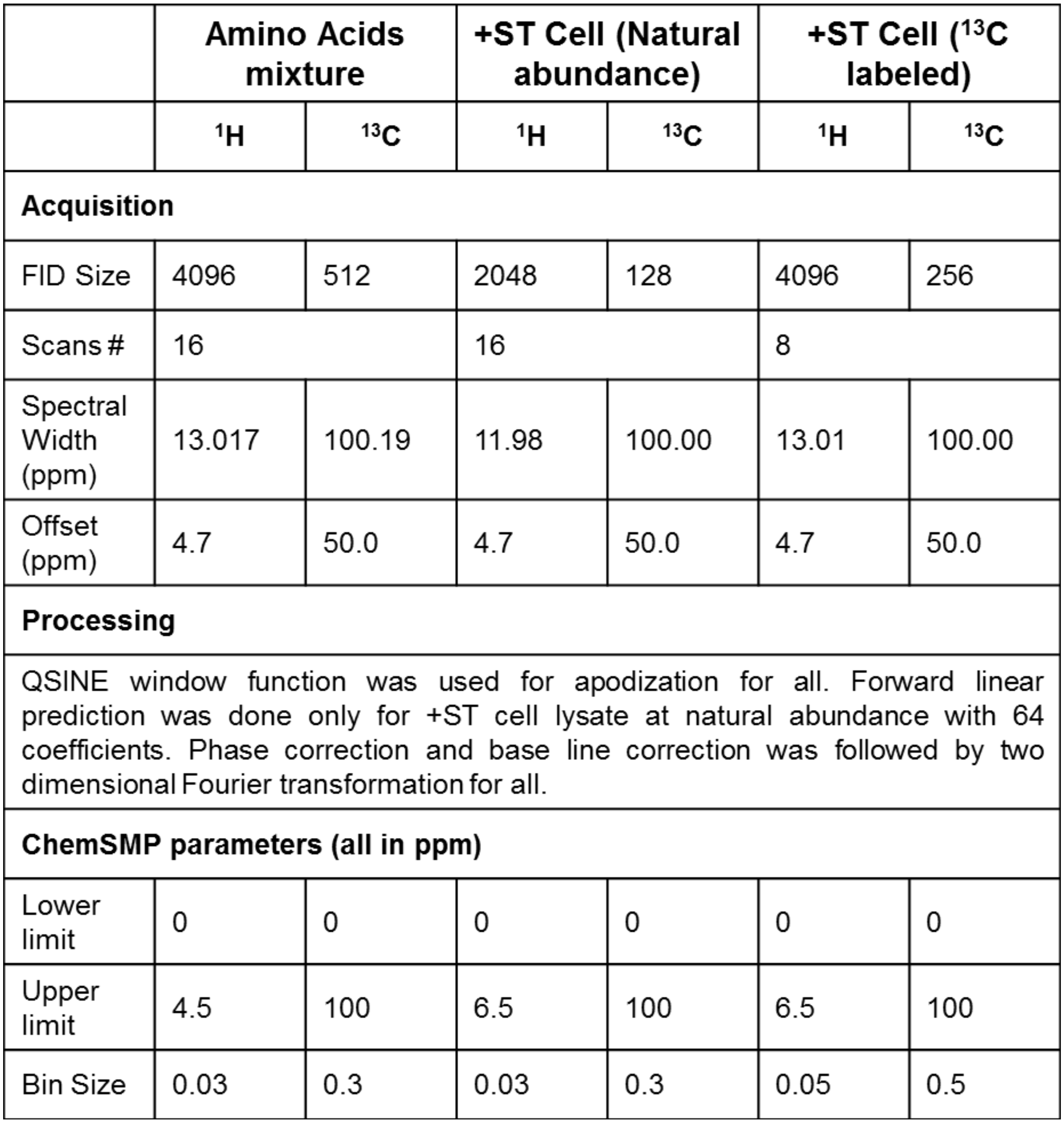
Sr	SMPDB ID	Metabolic Pathway	All		Chemical Shifts in Database	
			Total	**Unique**	**Total**	**Unique**
76	SMP00457	Lactose Degradation	9	0	5	0
77	SMP00459	Pyruvaldehyde Degradation	9	0	3	0
78	SMP00462	Inositol Phosphate Metabolism	25	3	4	0
79	SMP00463	Phosphatidylinositol Phosphate Metabolism	19	9	3	0
80	SMP00464	Vitamin K Metabolism	13	5	3	0
81	SMP00465	Carnitine Synthesis	20	4	10	0
82	SMP00466	Transfer of Acetyl Groups into Mitochondria	22	0	12	0
83	SMP00467	Trehalose Degradation	10	1	4	1
84	SMP00468	Degradation of Superoxides	10	1	1	0
85	SMP00479	Plasmalogen Synthesis	18	10	4	1
86	SMP00480	Mitochondrial Beta-Oxidation of Short Chain Saturated Fatty Acids	19	0	6	0
87	SMP00481	Mitochondrial Beta-Oxidation of Medium Chain Saturated Fatty Acids	25	0	6	0
88	SMP00482	Mitochondrial Beta-Oxidation of Long Chain Saturated Fatty Acids	27	5	8	1
89	SMP00654	Warburg Effect	60	1	30	0
90	SMP00715	Methylhistidine Metabolism	4	0	3	0
91	SMP00716	Thyroid hormone synthesis	13	6	5	4

Table S2: List of amino acids in the sample of amino acid mixtures along with their HMDB IDs



HMDB ID	Amino Acid
HMDB00161	L-Alanine
HMDB00574	L-Cysteine
HMDB00191	L-Aspartic acid
HMDB00148	L-Glutamic acid
HMDB00159	L-Phenylalanine
HMDB00123	Glycine
HMDB00177	L-Histidine
HMDB00172	L-Isoleucine
HMDB00182	L-Lysine
HMDB00687	L-Leucine
HMDB00696	L-Methionine
HMDB00168	L-Asparagine
HMDB00162	L-Proline
HMDB00641	L-Glutamine
HMDB00517	L-Arginine
HMDB00187	L-Serine
HMDB00167	L-Threonine
HMDB00883	L-Valine
HMDB00929	L-Tryptophan
HMDB00158	L-Tyrosine

Table S3: Parameters for three samples for acqusition of 2D [^{13}C, ^{1}H] HSQC NMR experiment and metabolic pathway analysis using ChemSMP.



	Amino Acids mixture		**+ST Cell (Natural abundance)**		**+ST Cell (^{13}C labeled)**	
	^{1}H	**^{13}C**	**^{1}H**	**^{13}C**	**^{1}H**	**^{13}C**
Acquisition						
FID Size	4096	512	2048	128	4096	256
Scans #	16		16		8	
Spectral Width (ppm)	13.017	100.19	11.98	100.00	13.01	100.00
Offset (ppm)	4.7	50.0	4.7	50.0	4.7	50.0
Processing						
QSINE window function was used for apodization for all. Forward linear prediction was done only for +ST cell lysate at natural abundance with 64 coefficients. Phase correction and base line correction was followed by two dimensional Fourier transformation for all.						
ChemSMP parameters (all in ppm)						
Lower limit	0	0	0	0	0	0
Upper limit	4.5	100	6.5	100	6.5	100
Bin Size	0.03	0.3	0.03	0.3	0.05	0.5

Table S4: Metabolic pathways (present in SMPDB) obtained as result on giving 20 amino acid names as input to MetPA. All the pathways except those marked in 'red' are common in the result obtained using ChemSMP and MetPA. Pathways 'SMP00019' and 'SMP00032' were absent in the pathways data used by ChemSMP. Pathways 'SMP00065' and 'SMP00076' do not have the 20 amino acids in them as per the SMPDB.

Sr.	SMPDB Pathway
1.	SMP00004
2.	SMP00006
3.	SMP00008
4.	SMP00009
5.	SMP00013
6.	SMP00015
7.	SMP00016
8.	SMP00019
9.	SMP00020
10.	SMP00021
11.	SMP00027
12.	SMP00029
13.	SMP00032
14.	SMP00033
15.	SMP00034
16.	SMP00035
17.	SMP00037
18.	SMP00041
19.	SMP00044
20.	SMP00046
21.	SMP00048
22.	SMP00050
23.	SMP00055
24.	SMP00063
25.	SMP00065
26.	SMP00066
27.	SMP00067
28.	SMP00072
29.	SMP00073
30.	SMP00076

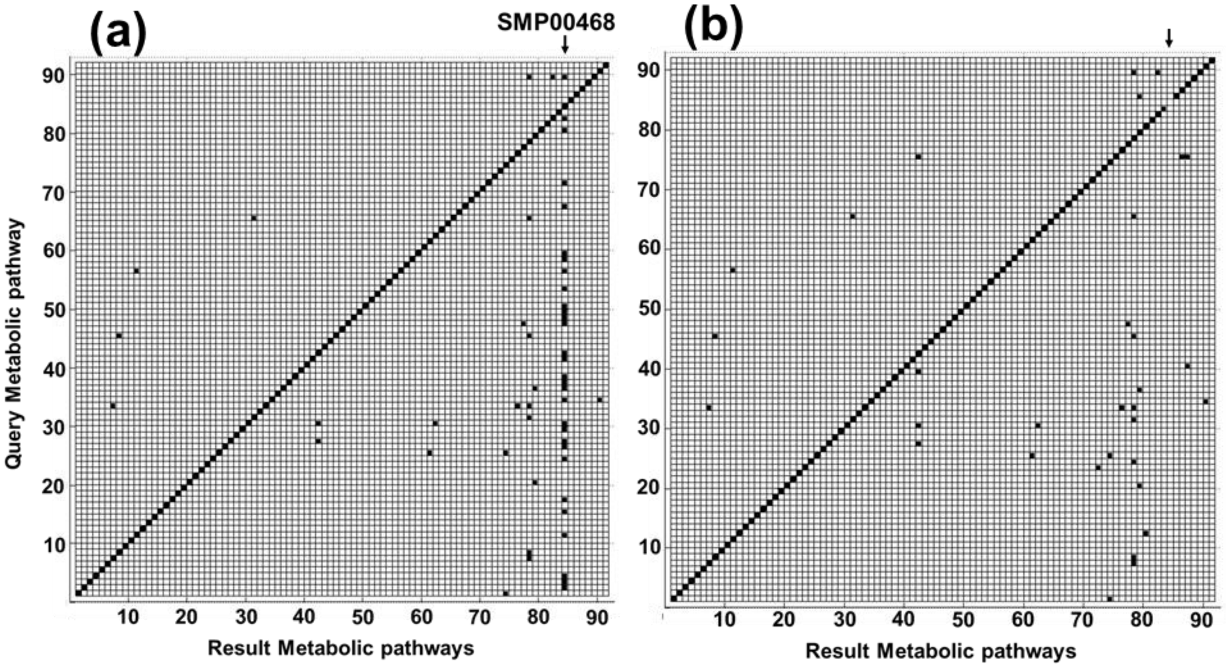


Figure S1: Results of simulation when chemical shifts of a single pathway was given as input and searched against the other 90 metabolic pathways. The diagonal shows all 91 metabolic pathways were scored with 100% coverage score. The off diagonal elements are the metabolic pathways which got 100% coverage score when chemical shifts from different metabolic pathway were queried. (a) An example pathway, SMP00468 is marked which has a single metabolite's chemical shifts in database and that metabolite is shared by several metabolic pathways. (b) Result obtained for similar analysis as in (a) except that a common metabolite NADP is neglected. SMP00468 no longer appears as false positive. Though, few new false positive cases appear because earlier some pathways were distinguished based on the metabolite NADP.

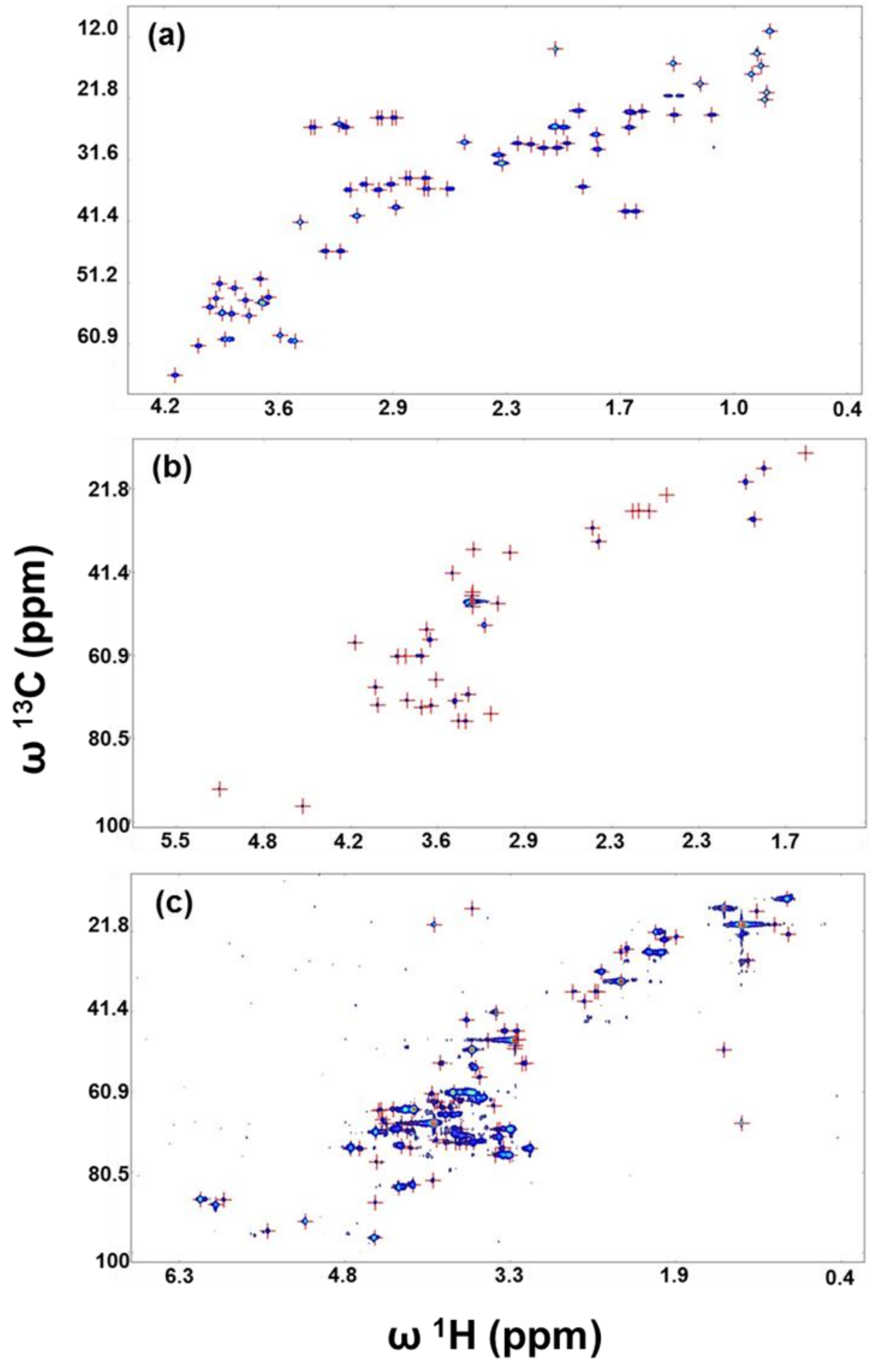
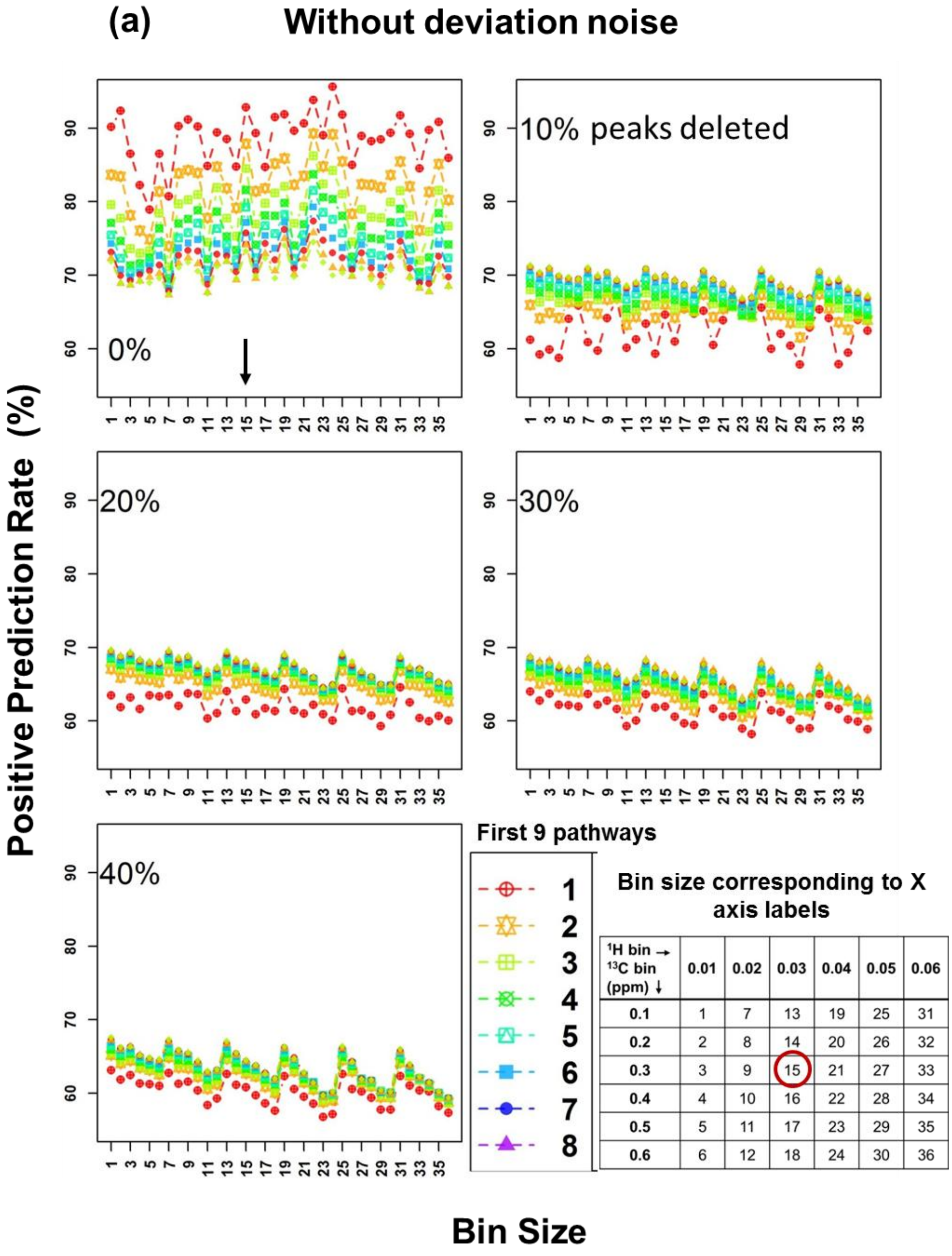


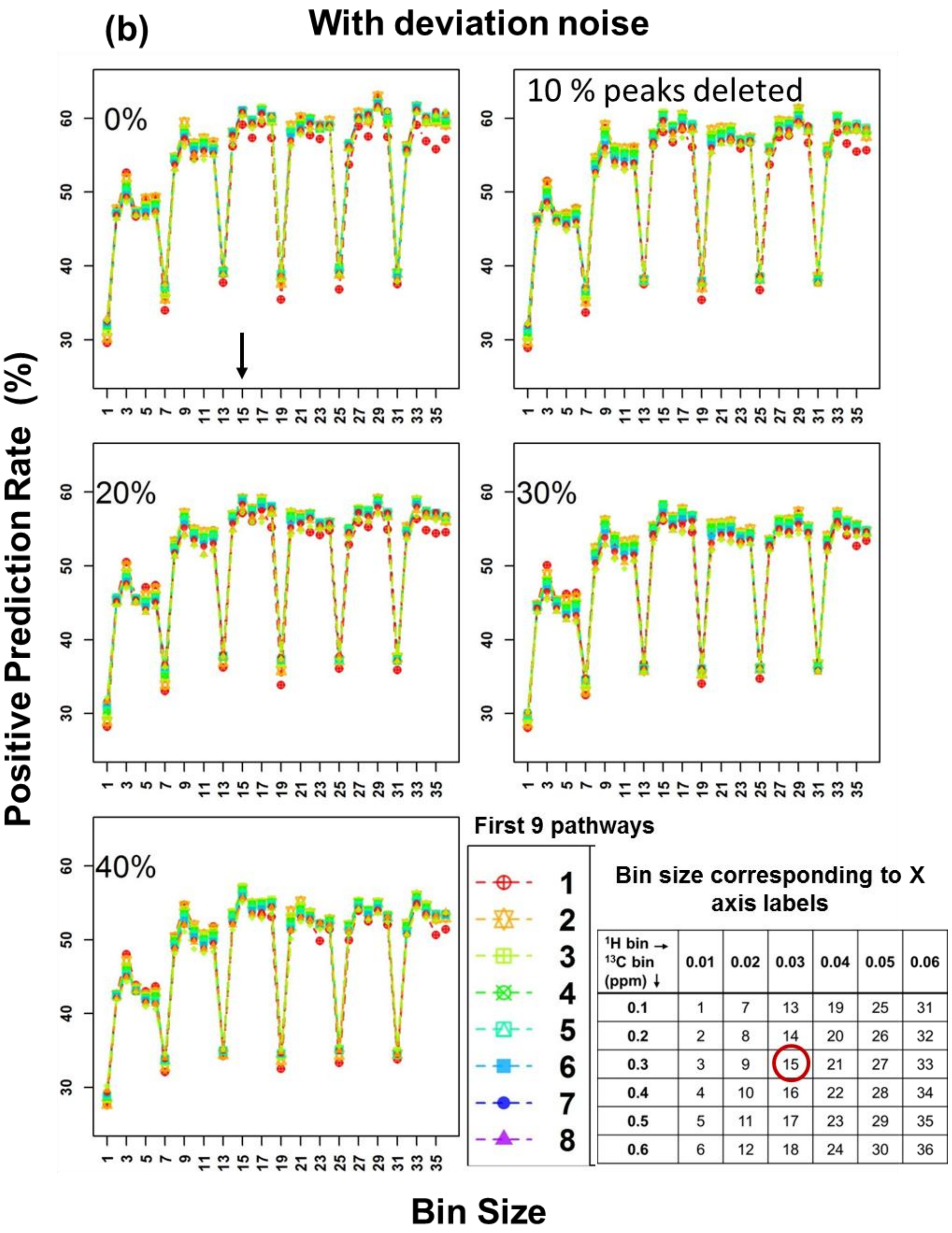
Figure S2: ChemSMP can directly take the processed NMR spectrum as input and can do automated peak picking using scripts provided in ChemSMP and 'nmrglue' package. A Bruker 2D [^{13}C, ^{1}H] HSQC spectrum used as input is shown in (a) mixture of 20 amino acids at natural abundance, (b) hf +ST cell

lysate at natural abundance and (c) hf +ST cell lysate using uniformly labeled ^{13}C glucose in growth media. The peaks picked automatically are marked in the spectral contours are shown in blue.



Bin size corresponding to X axis labels

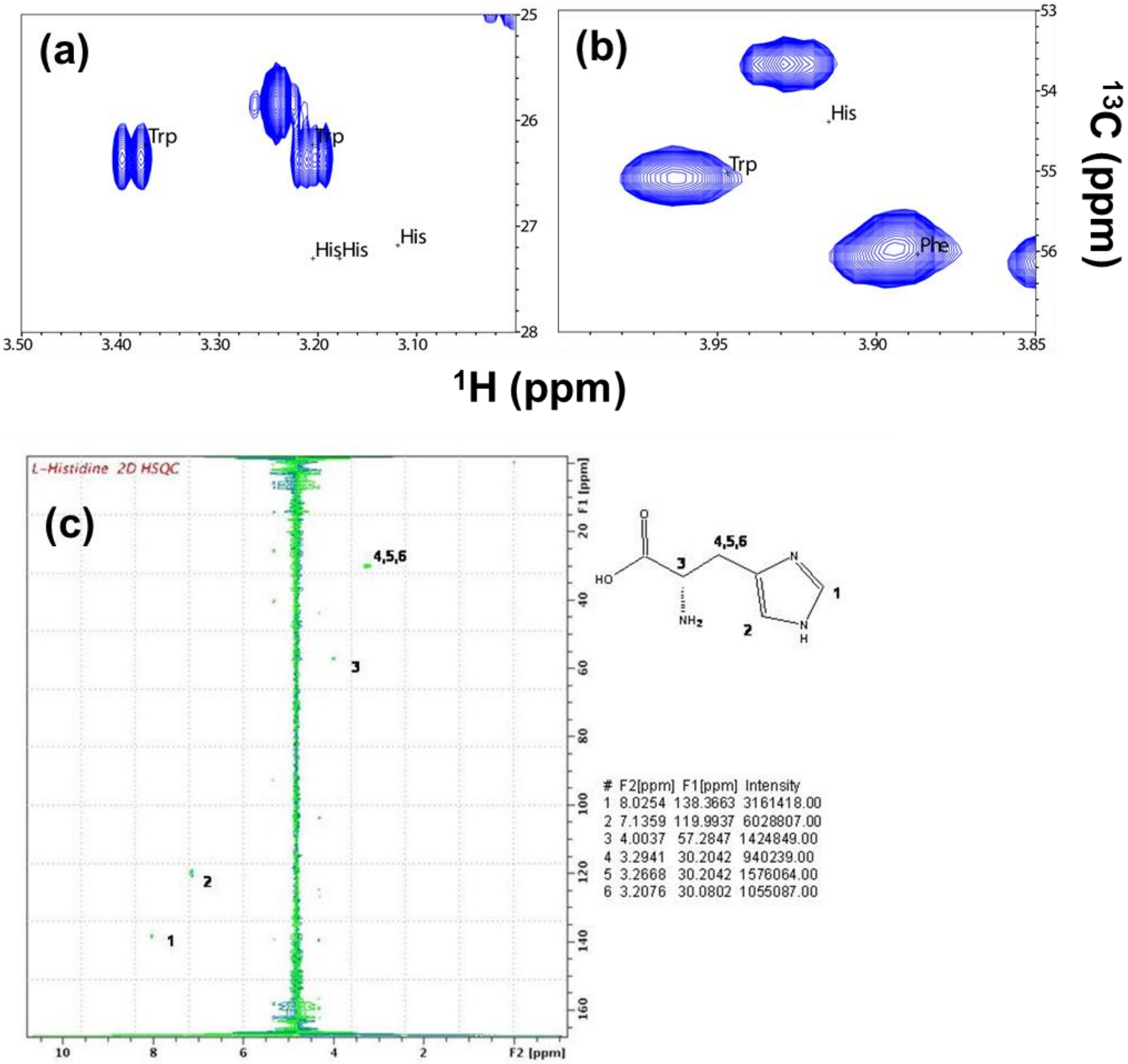
^{1}H bin → ^{13}C bin (ppm) ↓	0.01	0.02	0.03	0.04	0.05	0.06
0.1	1	7	13	19	25	31
0.2	2	8	14	20	26	32
0.3	3	9	15	21	27	33
0.4	4	10	16	22	28	34
0.5	5	11	17	23	29	35
0.6	6	12	18	24	30	36



^1H bin → ^{13}C bin (ppm) ↓	0.01	0.02	0.03	0.04	0.05	0.06
0.1	1	7	13	19	25	31
0.2	2	8	14	20	26	32
0.3	3	9	15	21	27	33
0.4	4	10	16	22	28	34
0.5	5	11	17	23	29	35
0.6	6	12	18	24	30	36

Figure S3: Evaluating the performance of ChemSMP on simulated dataset by varying the bin size in ^1H and ^{13}C dimensions under various conditions: (a) with peaks randomly deleted from 0% - 40% but without any chemical shift deviations added and (b) with peaks randomly deleted from 0% - 40% and with random chemical shift deviation (0.005-0.015 ppm in ^1H and 0.05-0.15 ppm in ^{13}C) added. Position

marked with an arrow and circled denotes the bin size used in this study i.e., 0.03 ppm in ^{1}H and 0.3 ppm in ^{13}C dimension. If the bin size is less than the deviations we see significant drop in Positive Prediction Rate (PPR) percentage. This can be seen in (b) at x = 1,7,13, 19, 25 and 31. At these positions the bin size in ^{1}H dimension is 0.01 ppm which is on an average half the times less than the noise (random number between 0.005 and 0.015) added to chemical shifts. We also see drop in PPR% for lower positions (position 1 with red curve) on increasing the bin size. This owes to decrease in specificity. This drop is less dramatic compared to if we use bin size less than the deviations expected in chemical shifts.



#	F2[ppm]	F1[ppm]	Intensity
1	8.0254	138.3663	3161418.00
2	7.1359	119.9937	6028807.00
3	4.0037	57.2847	1424849.00
4	3.2941	30.2042	940239.00
5	3.2668	30.2042	1576064.00
6	3.2076	30.0802	1055087.00

Figure S4: Expanded regions of 2D [^{13}C, ^{1}H] HSQC spectrum of 20 amino acids. After calibrating HMDB chemical shifts for large deviations Trp matches with peaks but His does not match for (a) chemical shift of β carbon and (b) chemical shift of α carbon. (c) 2D [^{13}C, ^{1}H] HSQC spectra and assignment image of L-Histidine downloaded from the HMDB database. It shows three chemical shifts for β carbon of Histidine.