Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information for

Accurate and Cost-Effective NMR Chemical Shift Predictions for Proteins using Molecules-in-Molecules Fragmentation-Based Method

*Sruthy Kettidathil Chandy, Bishnu Thapa, and Krishnan Raghavachari**

Department of Chemistry, Indiana University, Bloomington, Indiana 47405 United States

Table S1. Relative energies and percentage abundance of all conformers of 2LHY protein calculated using B3LYP-D3BJ/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod9	−4217.605132	0.00	1.00
2	mod2	−4217.598697	4.04	0.00
3	mod1	−4217.597872	4.56	0.00
4	mod7	−4217.595373	6.12	0.00
5	mod29	−4217.590107	9.43	0.00
6	mod4	−4217.583909	13.32	0.00
7	mod8	−4217.571801	20.92	0.00
8	mod23	−4217.571771	20.93	0.00
9	mod5	−4217.571150	21.32	0.00
10	mod15	−4217.565858	24.64	0.00
11	mod14	−4217.563936	25.85	0.00
12	mod26	−4217.560479	28.02	0.00
13	mod6	−4217.560119	28.25	0.00
14	mod12	−4217.554163	31.98	0.00
15	mod18	−4217.548329	35.64	0.00
16	mod24	−4217.546121	37.03	0.00
17	mod22	−4217.545765	37.25	0.00
18	mod27	−4217.544334	38.15	0.00
19	mod3	−4217.544174	38.25	0.00
20	mod16	−4217.544029	38.34	0.00
21	mod17	−4217.542834	39.09	0.00
22	mod28	−4217.530542	46.81	0.00
23	mod20	−4217.520019	53.41	0.00
24	mod21	−4217.518396	54.43	0.00
25	mod13	−4217.513346	57.60	0.00
26	mod10	−4217.504201	63.34	0.00
27	mod19	−4217.502760	64.24	0.00
28	mod25	−4217.491924	71.04	0.00

Table S2. Relative energies and percentage abundance of all conformers of 2LHY protein calculated using CAM-B3LYP-D3BJ/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod9	−4215.557636	0.00	1.00
2	mod2	−4215.549545	5.08	0.00
3	mod1	−4215.549260	5.26	0.00
4	mod7	−4215.547647	6.27	0.00
5	mod29	−4215.541881	9.89	0.00
6	mod4	−4215.535614	13.82	0.00
7	mod23	−4215.523978	21.12	0.00
8	mod5	−4215.523248	21.58	0.00
9	mod8	−4215.522859	21.82	0.00
10	mod15	−4215.517198	25.38	0.00
11	mod14	−4215.515322	26.55	0.00
12	mod26	−4215.512557	28.29	0.00
13	mod6	−4215.512199	28.51	0.00
14	mod12	−4215.505207	32.90	0.00
15	mod18	−4215.500275	35.99	0.00
16	mod24	−4215.497590	37.68	0.00
17	mod22	−4215.496909	38.11	0.00
18	mod3	−4215.496244	38.52	0.00
19	mod27	−4215.496161	38.58	0.00
20	mod16	−4215.495662	38.89	0.00
21	mod17	−4215.495530	38.97	0.00
22	mod28	−4215.483570	46.48	0.00
23	mod20	−4215.471325	54.16	0.00
24	mod21	−4215.470594	54.62	0.00
25	mod13	−4215.464809	58.25	0.00
26	mod10	−4215.456580	63.41	0.00
27	mod19	−4215.453852	65.13	0.00
28	mod25	−4215.443677	71.51	0.00

Table S3. Relative energies and percentage abundance of all conformers of 2LHY protein using *m*PW1PW91/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod9	−4216.247940	0.00	1.00
2	mod1	−4216.235269	7.95	0.00
3	mod2	−4216.233391	9.13	0.00
4	mod7	−4216.224658	14.61	0.00
5	mod29	−4216.221633	16.51	0.00
6	mod4	−4216.210082	23.76	0.00
7	mod23	−4216.209965	23.83	0.00
8	mod8	−4216.205740	26.48	0.00
9	mod5	−4216.204428	27.30	0.00
10	mod15	−4216.191910	35.16	0.00
11	mod14	−4216.191757	35.26	0.00
12	mod26	−4216.189997	36.36	0.00
13	mod6	−4216.187805	37.73	0.00
14	mod18	−4216.184597	39.75	0.00
15	mod12	−4216.181736	41.54	0.00
16	mod22	−4216.178705	43.45	0.00
17	mod24	−4216.175452	45.49	0.00
18	mod17	−4216.174407	46.14	0.00
19	mod16	−4216.174128	46.32	0.00
20	mod27	−4216.174006	46.39	0.00
21	mod3	−4216.168003	50.16	0.00
22	mod28	−4216.156501	57.38	0.00
23	mod21	−4216.153161	59.47	0.00
24	mod20	−4216.145957	64.00	0.00
25	mod13	−4216.142305	66.29	0.00
26	mod10	−4216.137874	69.07	0.00
27	mod19	−4216.127212	75.76	0.00
28	mod25	−4216.125100	77.08	0.00

Table S4. Relative energies and percentage abundance of all conformers of 2LHY protein using ωB97X-D/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod9	−4216.062582	0.00	1.00
2	mod2	−4216.056811	3.62	0.00
3	mod1	−4216.056332	3.92	0.00
4	mod7	−4216.054001	5.38	0.00
5	mod29	−4216.047825	9.26	0.00
6	mod4	−4216.042143	12.83	0.00
7	mod23	−4216.030499	20.13	0.00
8	mod8	−4216.029431	20.80	0.00
9	mod5	−4216.029022	21.06	0.00
10	mod15	−4216.023065	24.80	0.00
11	mod14	−4216.021646	25.69	0.00
12	mod6	−4216.018683	27.55	0.00
13	mod26	−4216.017819	28.09	0.00
14	mod12	−4216.011450	32.09	0.00
15	mod18	−4216.005467	35.84	0.00
16	mod24	−4216.004362	36.53	0.00
17	mod16	−4216.002687	37.58	0.00
18	mod27	−4216.002598	37.64	0.00
19	mod22	−4216.002450	37.73	0.00
20	mod3	−4216.001845	38.11	0.00
21	mod17	−4215.999041	39.87	0.00
22	mod28	−4215.989681	45.75	0.00
23	mod20	−4215.979290	52.27	0.00
24	mod21	−4215.975258	54.80	0.00
25	mod13	−4215.969246	58.57	0.00
26	mod10	−4215.962700	62.68	0.00
27	mod19	−4215.959747	64.53	0.00
28	mod25	−4215.947899	71.96	0.00

Table S5. Relative energies of all conformers of 2LI1 protein calculated using B3LYP-D3BJ/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t the lowest energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod19	−3473.957552	0.00	53.55
2	mod23	−3473.957384	0.11	44.82
3	mod1	−3473.954153	2.13	1.46
4	mod14	−3473.952064	3.44	0.16
5	mod18	−3473.949383	5.13	0.01
6	mod25	−3473.947594	6.25	0.00
7	mod5	−3473.941625	9.99	0.00
8	mod7	−3473.941583	10.02	0.00
9	mod8	−3473.939910	11.07	0.00
10	mod15	−3473.936815	13.01	0.00
11	mod3	−3473.936713	13.08	0.00
12	mod21	−3473.934775	14.29	0.00
13	mod22	−3473.934555	14.43	0.00
14	mod2	−3473.927605	18.79	0.00
15	mod24	−3473.922131	22.23	0.00
16	mod12	−3473.920116	23.49	0.00
17	mod17	−3473.919255	24.03	0.00
18	mod13	−3473.918946	24.23	0.00
19	mod9	−3473.917519	25.12	0.00
20	mod26	−3473.917424	25.18	0.00
21	mod27	−3473.912182	28.47	0.00
22	mod11	−3473.906185	32.23	0.00
23	mod20	−3473.902781	34.37	0.00
24	mod6	−3473.902028	34.84	0.00
25	mod10	−3473.902018	34.85	0.00
26	mod16	−3473.887233	44.13	0.00
27	mod4	−3473.885650	45.12	0.00

Table S6. Relative energies and percentage abundance of all conformers of 2LI1 protein calculated using CAM-B3LYP-D3BJ/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod19	−3472.240655	0.00	0.84
2	mod23	−3472.238987	1.05	0.14
3	mod1	−3472.237023	2.28	0.02
4	mod14	−3472.233625	4.41	0.00
5	mod18	−3472.232408	5.18	0.00
6	mod25	−3472.231410	5.80	0.00
7	mod5	−3472.224236	10.30	0.00
8	mod7	−3472.222637	11.31	0.00
9	mod8	−3472.221505	12.02	0.00
10	mod15	−3472.219289	13.41	0.00
11	mod3	−3472.218813	13.71	0.00
12	mod22	−3472.217601	14.47	0.00
13	mod21	−3472.217368	14.61	0.00
14	mod2	−3472.209683	19.44	0.00
15	mod24	−3472.206622	21.36	0.00
16	mod12	−3472.202298	24.07	0.00
17	mod13	−3472.202185	24.14	0.00
18	mod17	−3472.201671	24.46	0.00
19	mod9	−3472.201302	24.69	0.00
20	mod26	−3472.199241	25.99	0.00
21	mod27	−3472.195722	28.20	0.00
22	mod11	−3472.189389	32.17	0.00
23	mod20	−3472.186373	34.06	0.00
24	mod6	−3472.185465	34.63	0.00
25	mod10	−3472.184779	35.06	0.00
26	mod16	−3472.170305	44.15	0.00
27	mod4	−3472.168818	45.08	0.00

Table S7. Relative energies and percentage abundance of all conformers of 2LI1 protein calculated using *m*PW1PW91/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S. N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod23	−3472.836831	0.00	78.97
2	mod19	−3472.835579	0.79	20.96
3	mod1	−3472.830193	4.17	0.07
4	mod14	−3472.827205	6.04	0.00
5	mod18	−3472.826677	6.37	0.00
6	mod25	−3472.823056	8.64	0.00
7	mod5	−3472.816881	12.52	0.00
8	mod8	−3472.815964	13.09	0.00
9	mod7	−3472.811667	15.79	0.00
10	mod15	−3472.810528	16.51	0.00
11	mod22	−3472.809373	17.23	0.00
12	mod3	−3472.807330	18.51	0.00
13	mod21	−3472.803099	21.17	0.00
14	mod2	−3472.796484	25.32	0.00
15	mod17	−3472.794786	26.38	0.00
16	mod12	−3472.794776	26.39	0.00
17	mod24	−3472.794123	26.80	0.00
18	mod13	−3472.791070	28.72	0.00
19	mod9	−3472.790450	29.10	0.00
20	mod26	−3472.787739	30.81	0.00
21	mod27	−3472.787714	30.82	0.00
22	mod11	−3472.782273	34.24	0.00
23	mod10	−3472.779888	35.73	0.00
24	mod20	−3472.777952	36.95	0.00
25	mod6	−3472.776464	37.88	0.00
26	mod4	−3472.760723	47.76	0.00
27	mod16	−3472.759183	48.73	0.00

Table S8. Relative energies and percentage abundance of all conformers of 2LI1 protein calculated using ωB97X-D/6-31+G(d) level of theory. Conformers are ordered based on their relative energies.

S.N.	Conformers	Calculated Energies (Hartree)	Relative energy w.r.t lower energy conformer (kcal/mol)	Percentage abundance using Boltzmann weight
1	mod23	−3472.678534	0.00	58.82
2	mod19	−3472.678004	0.33	33.55
3	mod1	−3472.676531	1.26	7.04
4	mod14	−3472.674164	2.74	0.57
5	mod18	−3472.670348	5.14	0.01
6	mod25	−3472.668710	6.16	0.00
7	mod7	−3472.663125	9.67	0.00
8	mod5	−3472.662587	10.01	0.00
9	mod8	−3472.661628	10.61	0.00
10	mod15	−3472.658512	12.56	0.00
11	mod3	−3472.657923	12.93	0.00
12	mod21	−3472.656430	13.87	0.00
13	mod22	−3472.655888	14.21	0.00
14	mod2	−3472.649589	18.16	0.00
15	mod24	−3472.645137	20.96	0.00
16	mod17	−3472.642198	22.80	0.00
17	mod12	−3472.641236	23.40	0.00
18	mod13	−3472.640554	23.83	0.00
19	mod9	−3472.638485	25.13	0.00
20	mod26	−3472.638273	25.26	0.00
21	mod27	−3472.634757	27.47	0.00
22	mod11	−3472.628249	31.55	0.00
23	mod20	−3472.625916	33.02	0.00
24	mod6	−3472.623795	34.35	0.00
25	mod10	−3472.623233	34.70	0.00
26	mod4	−3472.608509	43.94	0.00
27	mod16	−3472.608123	44.18	0.00

Table S9. Mean absolute deviation (MAD) and Pearson correlation coefficient (R) between the experimental and calculated chemical shifts for various DFT methods. Vertical axis shows MAD values of ^{1}H, ^{13}C and ^{15}N NMR chemical shifts using MIM2[X/6-311++G(2d,2p):X/6-31G] of 2LI1 protein with respect to experimental values, where X represents different density functionals used for MIM2-NMR calculations. Horizontal axis shows different DFT methods used for calculation for Boltzmann weighted average with the 6-31+G(d) basis set.

BOLTZMANN / MIM2-NMR		B3LYP-D3BJ				CAM-B3LYP-D3BJ				WB97XD				MPW1PW91			
		^{1}H	^{13}C	^{15}N	total	^{1}H	^{13}C	^{15}N	total	^{1}H	^{13}C	^{15}N	total	^{1}H	^{13}C	^{15}N	total
MPW1PW91	MAD	0.78	2.42	7.25	1.80	0.76	2.74	5.51	1.67	0.80	2.36	8.10	1.70	0.84	2.36	19.31	2.27
	R	0.90	0.99	0.41		0.91	0.99	0.52		0.90	0.99	0.36		0.89	0.99	0.15	
CAM-B3LYP-D3BJ	MAD	0.81	2.62	7.17	1.74	0.80	2.95	5.48	1.76	0.82	2.54	7.99	1.76	0.86	2.47	19.79	2.34
	R	0.90	0.99	0.42		0.90	0.99	0.52		0.90	0.99	0.37		0.89	0.99	0.16	
B3LYP-D3BJ	MAD	0.80	2.54	10.82	1.89	0.78	2.84	5.51	1.72	0.81	2.57	8.62	1.80	0.84	2.56	24.39	2.60
	R	0.90	0.99	0.41		0.90	0.99	0.52		0.90	0.99	0.34		0.89	0.99	0.12	
WB97XD	MAD	0.78	2.49	7.12	1.68	0.77	2.81	5.38	1.69	0.79	2.40	8.03	1.70	0.82	2.33	18.82	2.23
	R	0.91	0.99	0.53		0.91	0.99	0.53		0.90	0.99	0.36		0.89	0.99	0.16	

Table S10. Mean absolute deviation (MAD) and Pearson correlation coefficient (R) calculated for the 2LI1 protein at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level of theory. The structures were minimized using molecular mechanics with various restraint parameters. (See **Methods** section of the main text for details)

Atoms	Statistics	Restraint parameter				
		0.25 Å	0.5 Å	1.0 Å	1.5 Å	2.0 Å
H	MAD	0.70	0.69	0.68	0.69	0.68
	R	0.91	0.92	0.92	0.92	0.92
C	MAD	2.62	2.48	2.70	2.17	2.41
	R	0.98	0.99	0.99	0.99	0.99
N	MAD	4.03	4.02	3.89	5.25	7.31
	R	0.61	0.60	0.62	0.56	0.46
Average	MAD	1.53	1.48	1.54	1.60	1.60

Table S11. Mean absolute deviation (MAD) and Pearson correlation coefficient (R) calculated for the 2LI1 protein at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level of theory. The structures were minimized using molecular mechanics with various restraint parameters. (See **Methods** section of the main text for details)

Atoms	Statistics	Restraint parameter				
		0.25 Å	0.5 Å	1.0 Å	1.5 Å	2.0 Å
H	MAD	0.66	0.64	0.69	0.69	0.70
	R	0.93	0.93	0.93	0.93	0.93
C	MAD	2.29	1.99	1.83	2.02	2.56
	R	0.99	0.99	0.99	0.99	0.99
N	MAD	3.38	4.01	3.65	3.27	4.38
	R	0.63	0.62	0.68	0.71	0.62
Total	MAD	1.35	1.26	1.22	1.27	1.49

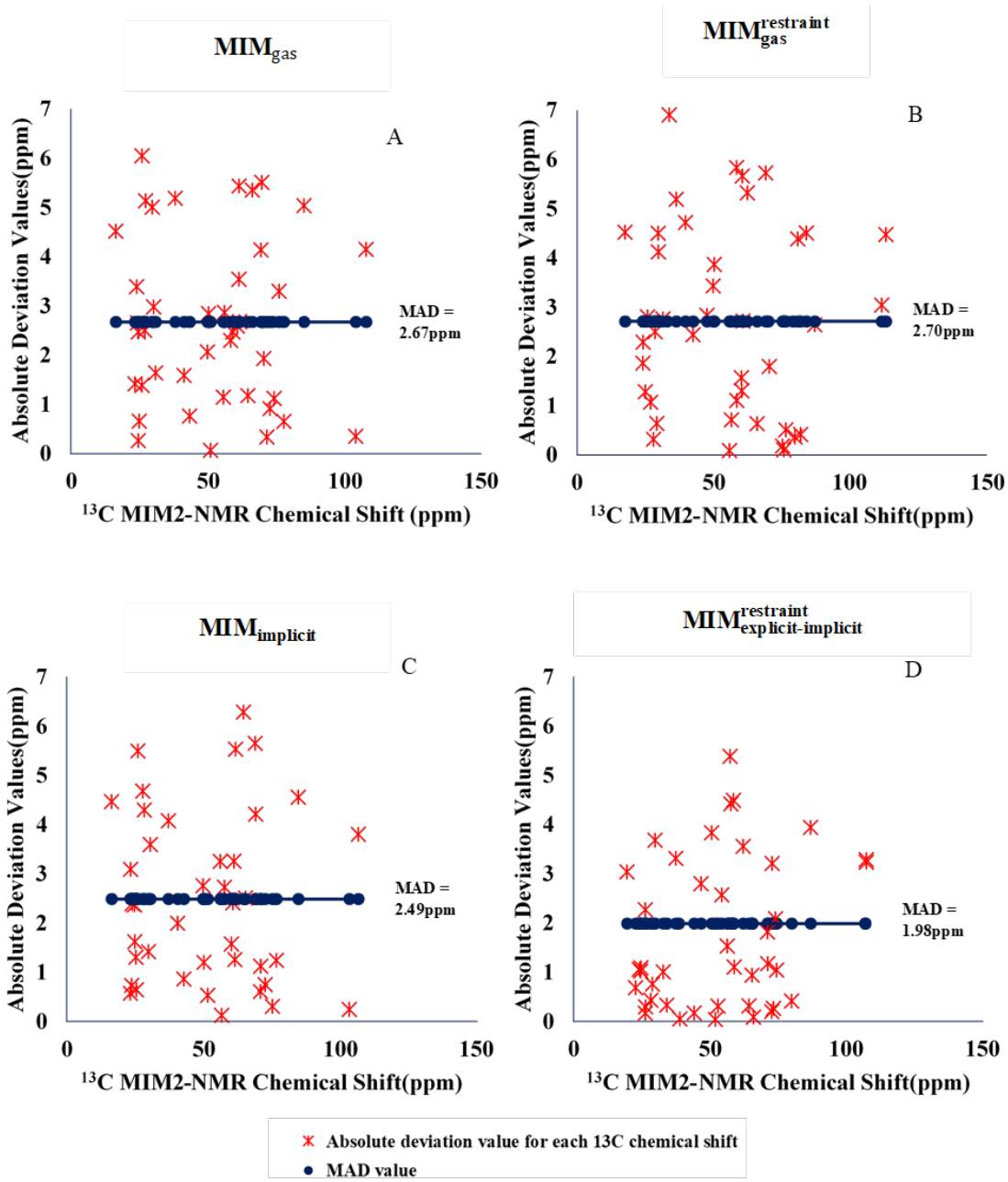

Figure S1. Distribution of absolute deviation values of ^{13}C MIM2-NMR chemical shifts between calculated and experimental chemical shifts of **2LHY** molecule. The MAD values for (A) MIM$_{gas}$, (B) MIM$^{restraint}_{gas}$, (C) MIM$_{implicit}$ and (D) MIM$^{restraint}_{explicit-implicit}$ models at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted with respect to tetramethylsilane (TMS) as reference.

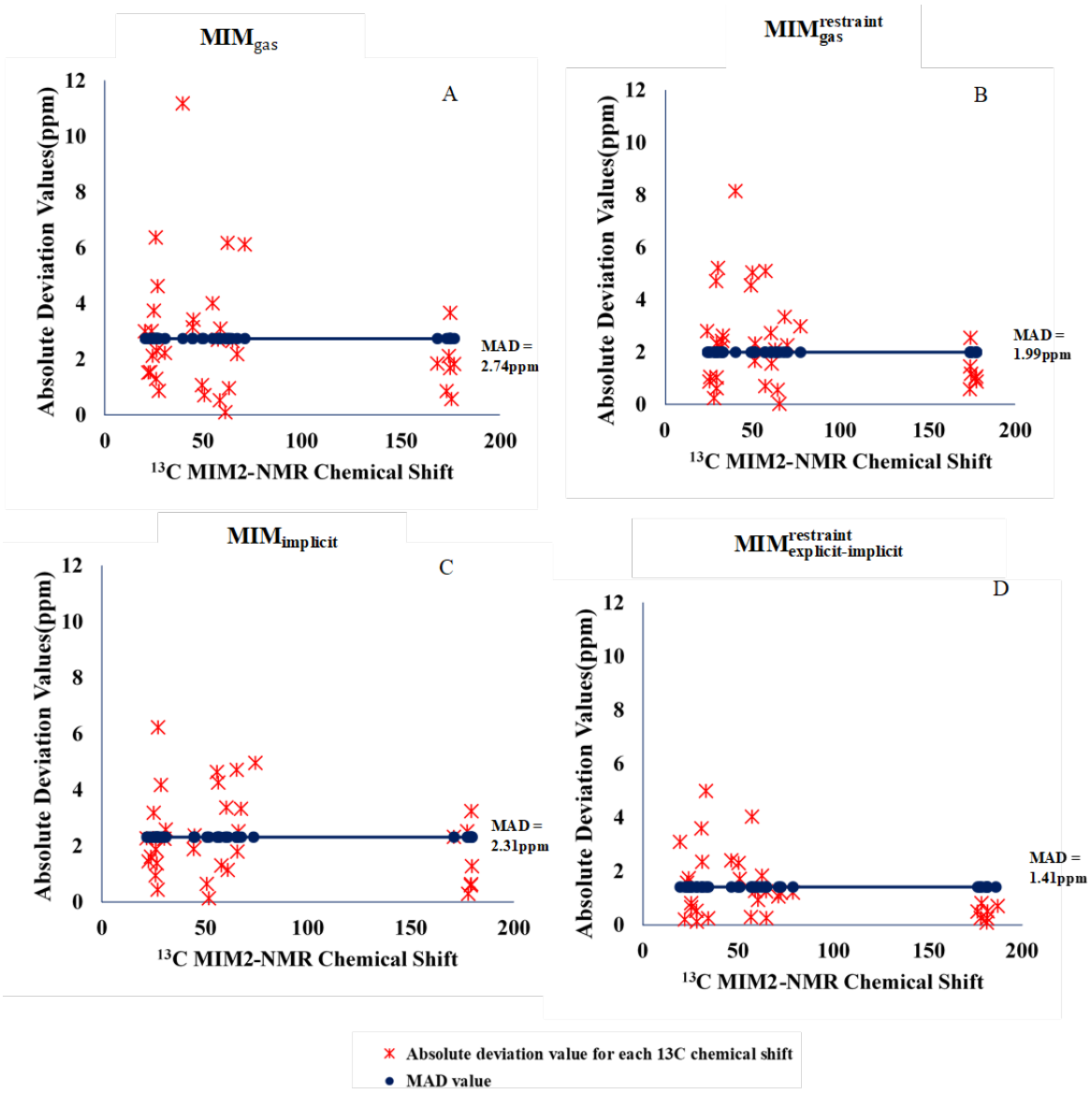

Figure S2. Distribution of absolute deviation values of ^{13}C MIM2-NMR chemical shifts between calculated and experimental chemical shifts of **2LI1** molecule. The MAD values for (A) MIM_{gas}, (B) $MIM^{restraint}_{gas}$, (C) $MIM_{implicit}$ and (D) $MIM^{restraint}_{explicit-implicit}$ models at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted with respect to tetramethylsilane (TMS) as reference.

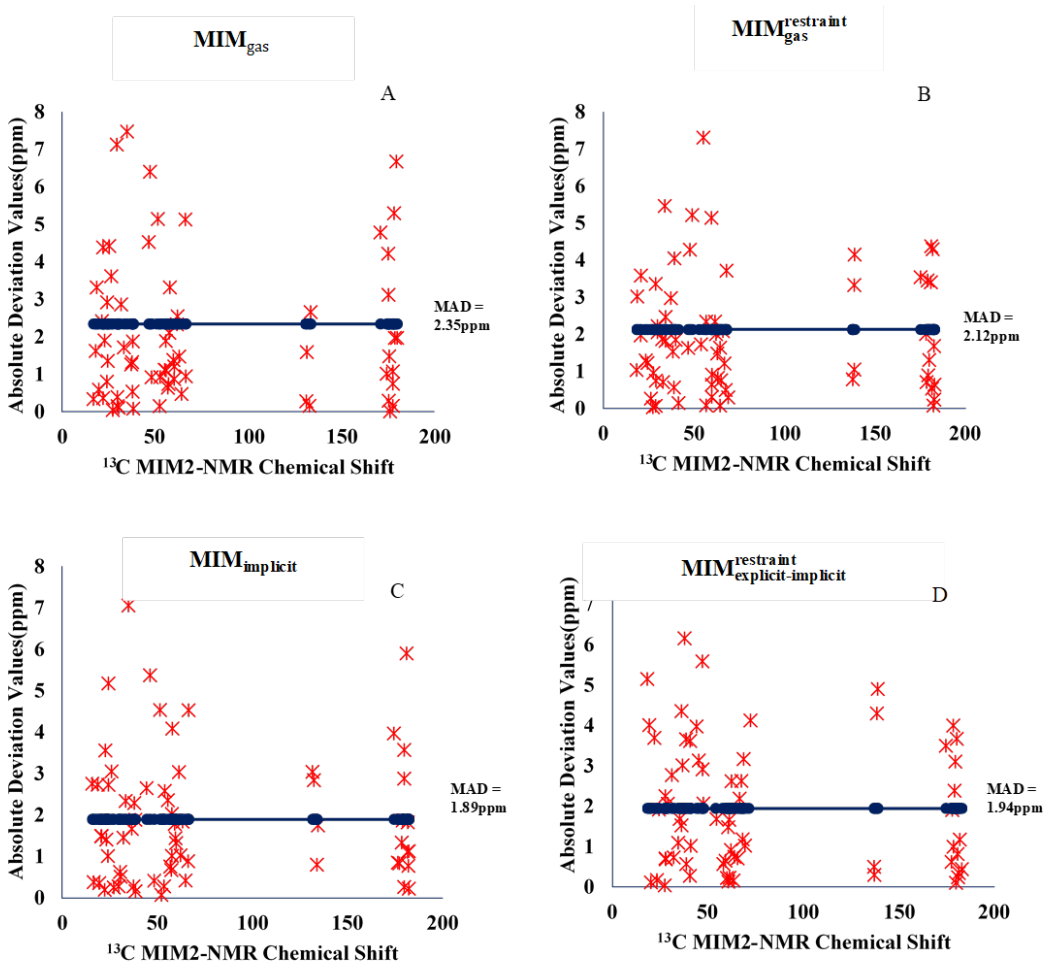

Figure S3. Distribution of absolute deviation values of ^{13}C MIM2-NMR chemical shifts between calculated and experimental chemical shifts of **2MC5** molecule. The MAD values for (A) MIM_{gas}, (B) $MIM^{restraint}_{gas}$, (C) $MIM_{implicit}$ and (D) $MIM^{restraint}_{explicit-implicit}$ models at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted with respect to tetramethylsilane (TMS) as reference.

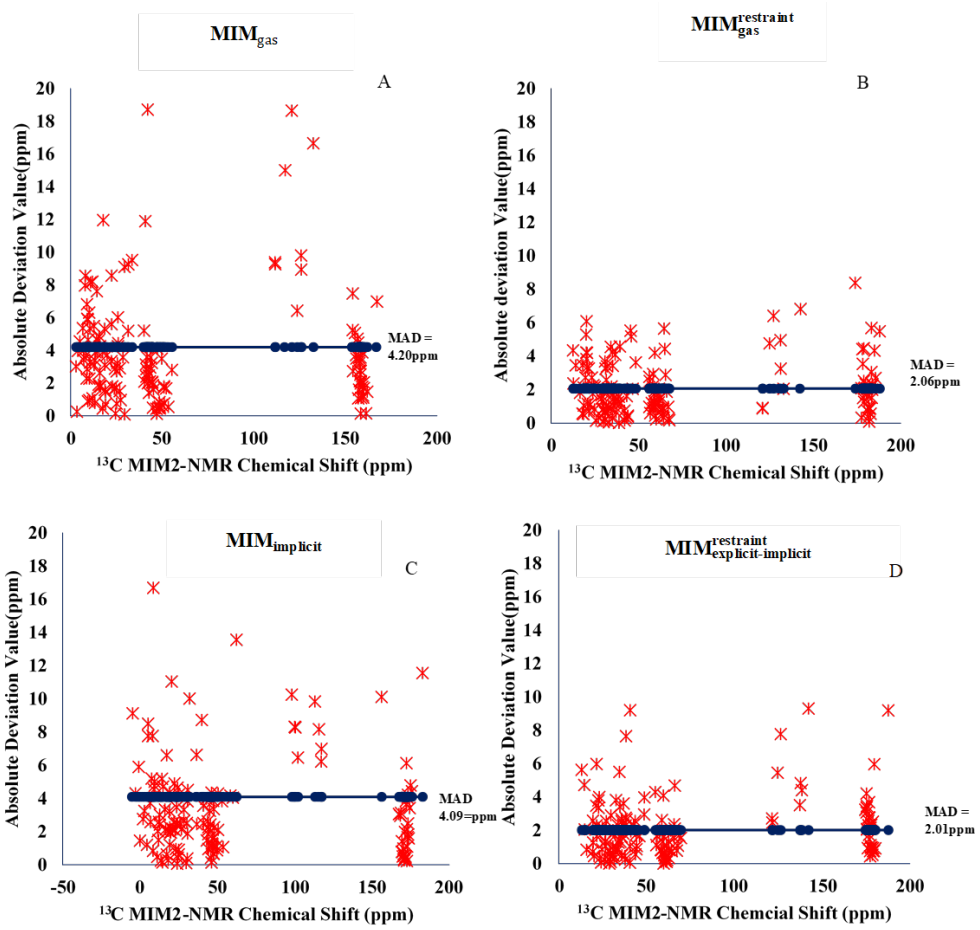

Figure S4. Distribution of absolute deviation values of ^{13}C MIM2-NMR chemical shifts between calculated and experimental chemical shifts of **3UMK** molecule. The MAD values for (A) MIM$_{gas}$, (B) MIM$^{restraint}_{gas}$, (C) MIM$_{implicit}$ and (D) MIM$^{restraint}_{explicit-implicit}$ models at MIM2[*m*PW1PW91/6-311++G(2d,2p):*m*PW1PW91/6-31G] level. The MIM ^{13}C NMR chemical shifts are depicted with respect to tetramethylsilane (TMS) as reference.

Table S12. Timings for a 33 peptide slice of 3UMK using 4-processor machine.

Structure	Number of water molecules	Natoms	Method	Type	NBasis	Timing (min)
3UMK	0	547	mPW1PW91/ 6-31G	MIM real low (unfragmented) MIM_{gas}	3000	5377
3UMK	0	547	mPW1PW91/ 6-31G	MIM real low (unfragmented) $MIM_{implicit}$	3000	7206
3UMK	3	90	mPW1PW91/ 6-311++G(2d,2p)	MIM model high (largest subsystem) $MIM^{restraint}_{explicit-implicit}$	1631	5514
3UMK	15	592	mPW1PW91/ 6-311++G(2d,2p)	MIM real high (unfragmented) $MIM^{restraint}_{explicit-implicit}$	10,790	Not Available

Note: XYZ coordinates of the structures studied can be found in: https://doi.org/10.5281/zenodo.4270566