SHIFTX2: Significantly Improved Protein Chemical Shift Prediction

Beomsoo Han[1], Yifeng Liu[1], Simon W. Ginzinger[4] and David S. Wishart[1,2,3†]

[1]Department of Computing Science, University of Alberta; [2]Department of Biological Sciences, University of Alberta; [3]National Research Council, National Institute for Nanotechnology (NINT), Edmonton, AB, Canada T6G 2E8 and [4]Department of Molecular Biology, Division of Bioinformatics, Center of Applied Molecular Engineering, University of Salzburg, Hellbrunnerstr. 34/3.OG, 5020 Salzburg, Austria

† To whom correspondence should be addressed. (Phone: 780-492-0383 email: david.wishart@ualberta.ca)

Supplementary Tables

Table S1. SHIFTX+ training dataset consisting of 197 PDB and BMRB pairs. The resolution (Res) of each structure is given on the right.[a]

PDB ID	BMRB #	Res. (Å)	PDB ID	BMRB #	Res. (Å)
3LZTA	4562	0.93	1LZ1	5142	1.50
1F94	5097	0.97	1U8TA	4472	1.50
1CEXA	4101	1.00	1VDQA	4562	1.50
1MN8D	5623	1.00	3EZMA	*cvn*	1.50
5PTI	*bpti*	1.00	1L2HA	1062	1.54
1CNR	6455, 6504	1.05	2GFEB	5182	1.54
1A6KA	4061	1.10	2A0B	4857	1.57
1D4TA	5211	1.10	1BFG	4091	1.60
1H4GA	5352	1.10	1BKFA	4077	1.60
1L3K	4084	1.10	1HCBA	4022	1.60
1U07	6375	1.13	1JL3B	6075	1.60
1RGEB	4259	1.15	1Q5PA	*maxacal*	1.60
1TJM	5194	1.18	1QRXA	*alpha_LP*	1.60
1M15A	6542	1.20	2AWG	6923	1.60
1NEY	7216	1.20	2D3D	6922	1.60
2FCLA	7086	1.20	2EWRA	7086	1.60
1RUV	4031	1.25	4FGFA	4091	1.60
1DBF	6494	1.30	4ICBA	390, 6699	1.60
1F4P	5571	1.30	2CPLA	*cyclophilin*	1.63
1PLC	4019	1.33	1F5RI	4968	1.65
1I1JB	5220	1.39	1G4CB	4299	1.65
1EW4	5792	1.40	1M45A	6332	1.65
1IWTA	5130, 5142	1.40	1NQDA	15722	1.65
1KJLA	4909	1.40	1SNCA	*snase*, 4052	1.65
1KQR	5275	1.40	2ITLA	4127	1.65
2CIAA	6575	1.45	2TRXA	62, 1812	1.68
2END	5244	1.45	1CLLA	547, 6023	1.70
2F3YA	5286, 6541, 15650, 15852	1.45	1DDW	4766	1.70
3RN3	4031	1.45	1EMV	4115	1.70
1M5E	5182	1.46	1ERTA	*thioredoxin_red*	1.70
2RN2	1657	1.48	1F2F	6503	1.70
8ABP	6136	1.49	1J54	6184	1.70
1A2PA	975, 4964	1.50	1MXEA	547, 1634	1.70
1F46A	4717	1.50	1ONC	4371	1.70
1HFCA	4064	1.50	1UB4B	6828, 6833	1.70
1HKA	4299	1.50	1W41A	5485	1.70
1ICMA	7356	1.50	1YKTB	5359	1.70
1J97	5799	1.50	1YPCI	4974	1.70

Table S1. (Continued)

PDB ID	BMRB #	Res. (Å)	PDB ID	BMRB #	Res. (Å)
2B8X	4094	1.70	1A7T	4102	1.85
2BKYA	5226	1.70	1AYF	4566	1.85
2BKYB	6019	1.70	1EB0A	5484, 5826	1.85
2F1YA	6939	1.70	1EPFC	4162	1.85
2GBTC	15712, 15713	1.70	1LAXA	4354	1.85
2GROA	4132	1.70	1ZNB	4102	1.85
2NNRA	6876	1.70	2GABB	5508	1.85
1FH9A	7264	1.72	3FAPB	6760	1.85
2FKEA	4077	1.72	1AILA	4317	1.90
1FE4B	6266	1.75	1B2VA	5081	1.90
1GNU	5058	1.75	1BRIB	4964	1.90
1JV4	4340	1.75	1EZ3	4198	1.90
1UJ8A	6776	1.75	1FF3A	6090	1.90
1TOPA	4401	1.78	1FF3C	6786	1.90
1UV0A	6231	1.78	1FQA	4354	1.90
1BDO	4425	1.80	1FZY	4567	1.90
1C44	4438	1.80	1G8I	4378	1.90
1CBS	4186	1.80	1H4H	5352	1.90
1CM2A	2371	1.80	1IAZA	4797	1.90
1EKG	4342	1.80	1IWMA	10096	1.90
1EXP_	7264	1.80	1NW2H	5241	1.90
1GZIA	15817	1.80	1Q4R	5843	1.90
1H70	6074	1.80	1QAV	6754	1.90
1HL5J	4202	1.80	1RMMA	5666	1.90
1IIBB	4955	1.80	1RSY	4039, 4041, 4167	1.90
1IU1	5761	1.80	1UP1_	4084	1.90
1NCX	5071, 5738	1.80	1VFQA	6283	1.90
1RX2_	4554, 5740	1.80	1Y2GB	4717	1.90
1SMXB	6122	1.80	1ZW9A	5355	1.90
1UBQA	*5387, ubiquitin*	1.80	2ADFA	5456	1.90
1UCKB	4259	1.80	2CDNA	4840	1.90
1XUO	4553	1.80	2D58	6980	1.90
1ZLQ	6416	1.80	1KBL	6932	1.94
2RNT	*Ribonouclease_T1*	1.80	1OSPO	4076	1.95
1BCX	4704, 4705	1.81	1P7TA	5471	1.95
1IV7A	4638, 5529	1.82	2VFX	5299	1.95
1ZE3C	4070	1.84	1H7M	5485	1.96
1ZE3D	6779	1.84	1A30A	*HIV protease (HIV)*	2.00

Table S1. (Continued)

PDB ID	BMRB #	Res. (Å)	PDB ID	BMRB #	Res. (Å)
1AM1	5355	2.00	1N2DA	6332	2.00
1C7FB	5540, 5571	2.00	1SCJA	15706	2.00
1CDLA	1634	2.00	1SN8B	6122	2.00
1FIL	4082	2.00	1TVQA	16310	2.00
1GXQ	4421	2.00	1TW4B	15084, 15854	2.00
1HPCB	4336	2.00	1U9B	4132	2.00
1IPB	5712	2.00	1UOH	5898	2.00
1MHOA	5206	2.00	1VC1	5921	2.00
1MJC	4296	2.00	1YP7A	4340	2.00
1MKAA	*dehydrase*	2.00	1ZJLA	4986	2.00
1N0S	5756	2.00	2BE6A	4174	2.00

[a]10 protein chemical shifts (marked in italics) were obtained from the TALOS database (Cornilescu et al. 1999), except Ribonouclease_T1 (Wishart et al. 1997))

Table S2: List of the 61 “test” proteins (PDB accession number, BMRB accession, resolution and size) used in the assessment of different shift prediction programs.

No	PDB	BMRB	Res. (Å)	Residues	No	PDB	BMRB	Res. (Å)	Residues
1	1KF3A	4032	1.05	124	32	1HUUB	4047	2.00	74
2	1XMTA	6338	1.15	95	33	1OQRC	4149	1.65	104
3	1LM4B	4834	1.45	183	34	1T3YA	6032	1.15	125
4	2B02A	6597	1.50	119	35	1O4DA	6604	1.85	102
5	2H30A	6709	1.60	143	36	1T8LB	5358	1.75	55
6	1TVGA	6344	1.60	136	37	1UDRD	4083	1.90	123
7	2D3GB	6457	1.70	72	38	1ODVA	6321	1.14	100
8	1SNMA	4053	1.74	136	39	1DQEA	6313	1.80	128
9	1JIWI	6292	1.74	105	40	1Y93A	6391	1.03	156
10	1O13A	6198	1.83	117	41	1W80A	6034	1.90	236
11	1T15A	6114	1.85	214	42	1V9TA	4037	1.70	165
12	2C5LC	6635	1.90	106	43	2AOJB	5967	1.60	95
13	1YSBB	6223	1.70	158	44	2ESPA	6277	1.52	147
14	1DYTA	15757	1.75	133	45	1TP5A	6193	1.54	105
15	1H4AX	16173	1.15	173	46	1TP9A	6132	1.62	162
16	1RROA	7322	1.30	108	47	2BF5B	4560	1.71	102
17	1RWYB	15517	1.05	109	48	1CWCA	2208	1.86	165
18	1T2WB	4879	1.80	142	49	1YKYX	4831	1.90	129
19	1VP6A	15249	1.70	133	50	1KDBA	6250	1.90	116
20	1YJ7A	6252	1.80	166	51	2A38C	5760	2.00	193
21	1YZ1C	15243	2.00	172	52	256BA	6560	1.40	106
22	1ZJLA	7114	2.00	368	53	1BT5A	6024	1.80	263
23	1ZX8A	16007	1.90	123	54	1SGZA	6016	2.00	341
24	2ES9A	15089	2.00	99	55	2IN0A	15560	1.60	139
25	2HZEA	4113	1.80	107	56	1SYDA	15232	1.70	136
26	2O0PA	15281	1.90	114	57	1JR2B	7242	1.84	260
27	2OYNA	15530	1.85	135	58	1U7BA	15501	1.88	255
28	2Z2IA	7055	1.98	179	59	2A0NA	15741	1.64	251
29	1HQ2A	4300	1.25	157	60	2DYIA	10139	2.00	162
30	1JTGC	6357	1.73	259	61	1B1HA	10053	1.80	517
31	1RUWA	6197	1.80	69					

Table S3. Features for building prediction models of SHIFTX+

Feature	Neighbor residues for i^{th} residue
Amino acid type	i-2, i-1, i, i+1, i+2
Secondary structure	i-1, i, i+1
Hydrogen bond status (yes/no)	i-1, i, i+1
Accessible Surface Area (ASA) (residual/fractional)	
Accessible Surface Area (ASA) (residual/fractional) for side chain	
Solvation Free Energy (SFE)	
Hydrogen bond lengths (HA1/HA2/HN/O)	i-1, i, i+1
Phi/Psi/Omega/Chi1 angles	i-2, i-1, i, i+1, i+2
Chi2/Chi3 angles	i
Theta, Kappa angles (H-bond angles)	i
Disulfide bond status (yes/no)	i-1, i, i+1
Hydrophobicity	
Atomic surface area (ASA) for each atom	
Random coil shift	
Ring current shift/effect	
Electric field shift contribution	
Hydrogen bond shift contribution	
pH	
Temperature (K)	

Table S4. 10-fold cross-validation and testing accuracy of SHIFTX+

Atom	No. of shifts	Cross Validation (10x)		Test dataset	
		Correlation	RMSD	Correlation	RMSD
^{15}N	25226	0.9149	2.2878	0.9063	2.3697
$^{13}C\alpha$	26314	0.9842	0.8743	0.9849	0.8467
$^{13}C\beta$	22428	0.9970	1.0099	0.9971	0.9676
$^{13}C'$	19486	0.8939	0.9945	0.8933	0.9483
^{1}HN	24761	0.8103	0.4356	0.8102	0.4112
$^{1}H\alpha$	18887	0.9226	0.2152	0.9219	0.2092

Table S5. 10-fold cross-validation and testing accuracy of SHIFTX+ for side chain nuclei (including HA2 and HA3 nuclei from glycine).

Atom[a]	No. of shifts	Cross Validation (10x)		Test dataset	
		Correlation	RMSD	Correlation	RMSD
CD	1892	0.9999	0.6912	0.9999	0.6253
CD1	2440	0.9996	1.3292	0.9997	1.2276
CD2	1472	0.9996	1.4436	0.9996	1.4171
CE	870	0.9966	0.7738	0.9987	0.4206
CE1	642	0.9929	0.9296	0.9900	1.0565
CE2	338	0.9958	0.6804	0.9900	0.9070
CG	4447	0.9998	0.8207	0.9995	0.7863
CG1	1577	0.9504	1.0303	0.9568	0.9657
CG2	2328	0.8777	1.0043	0.8779	1.0975
CZ	270	0.9949	1.2241	0.9932	1.2889
HA2	1410	0.7715	0.2501	0.7177	0.2446
HA3	1331	0.6367	0.2593	0.6468	0.2623
HB	3898	0.9858	0.1871	0.9837	0.1916
HB2	9346	0.9502	0.2346	0.9450	0.2416
HB3	8827	0.9450	0.2450	0.9378	0.2613
HD1	2953	0.9974	0.2152	0.9976	0.2102
HD2	3479	0.9956	0.2075	0.9949	0.2426
HD3	1603	0.9688	0.2280	0.9747	0.2079
HE	341	0.9935	0.3115	0.9882	0.3998
HE1	1110	0.9529	0.3690	0.9639	0.3013
HE2	1338	0.9957	0.1835	0.9959	0.1812
HE3	776	0.9907	0.2141	0.9945	0.1724
HG	1039	0.6434	0.2483	0.6235	0.2441
HG1	1034	0.6944	0.1687	0.7071	0.1506
HG12	723	0.5249	0.3412	0.4421	0.3644
HG13	649	0.4471	0.3585	0.5000	0.3669
HG2	6166	0.9624	0.1761	0.9656	0.1725
HG3	3173	0.9067	0.2034	0.9160	0.2008
HZ	337	0.6693	0.2741	0.5862	0.3312

Table S6. SHIFTY+ performance for the complete set of 235 non-redundant proteins in the training and test dataset.

Atom	Correlation	RMSD	No. of predicted PDBs	Coverage
^{15}N	0.9800	1.1352	175	74.5%
^{13}Ca	0.9925	0.6127	170	72.3%
^{13}Cb	0.9991	0.5562	159	67.7%
^{13}C’	0.9638	0.5784	120	51.1%
^{1}HN	0.9610	0.2097	175	74.5%
^{1}Ha	0.9677	0.1411	139	59.1%
HB	0.9969	0.0877	124	52.8%
HB2	0.9809	0.1468	142	60.4%
HB3	0.9829	0.1402	126	53.6%
HD2	0.9956	0.1105	117	49.8%
HD3	0.9885	0.1439	111	47.2%
HE2	0.9150	0.0820	110	46.8%
HE3	0.9374	0.0808	89	37.9%
HG2	0.9546	0.2098	120	51.1%
HG3	0.8993	0.2324	119	50.6%
All Atoms			179	76.2%

Table S7. SHIFTY+ performance for the 1903 non-redundant proteins in RefDB.

Atom	Correlation	RMSD	No. of predicted PDBs	Coverage
^{15}N	0.9564	1.5721	1106	58.1%
^{13}Ca	0.9946	0.5057	1027	54.0%
^{13}Cb	0.9992	0.5147	978	51.4%
^{13}C’	0.9585	0.6076	786	41.3%
^{1}HN	0.9494	0.2133	1249	65.6%
^{1}Ha	0.9601	0.1440	1131	59.4%
HB	0.9926	0.1655	1073	56.4%
HB2	0.9554	0.2314	1128	59.3%
HB3	0.9550	0.2301	1079	56.7%
HD2	0.9962	0.1593	1055	55.4%
HD3	0.9848	0.1656	999	52.5%
HE2	0.9978	0.0000	960	50.4%
HE3	0.9915	0.1195	880	46.2%
HG2	0.9508	0.2096	1060	55.7%
HG3	0.9591	0.1416	1058	55.6%
All Atoms			1270	66.7%

Table S8. Performance comparison for SHIFTY+ with different sizes of the RefDB using the 235 proteins in training/test dataset as the query set.

Size of RefDB	Average coverage rate for backbone atoms	Average coverage rate for 15 predictable atoms
400	30.0	20.7
800	42.1	34.6
1200	53.9	42.6
1600	60.6	46.1
2000	64.5	51.4
< 3000	67.0	57.1

Table S9. Performance comparison for SHIFTX+, SHIFTY+, and SHIFTX2. The accuracies for SHIFTX+ and SHIFTX2 were calculated using all 61 proteins in the test dataset; statistics for SHIFTY+ were calculated using 46 proteins with SHIFTY+ predictions.

Atom	No. of shifts	Correlation			RMSD		
		SHIFTX+	SHIFTY+	SHIFTX2	SHIFTX+	SHIFTY+	SHIFTX2
^{15}N	8135	0.9044	0.9974	0.9800	2.3928	0.4115	1.1169
^{13}Cα	8704	0.9833	0.9991	0.9959	0.8891	0.2087	0.4412
^{13}Cβ	7431	0.9970	0.9999	0.9992	0.9849	0.2136	0.5163
^{13}C’	7190	0.8890	0.9961	0.9676	0.9661	0.1847	0.5330
^{1}HN	8098	0.7939	0.9964	0.9714	0.4375	0.0630	0.1711
^{1}Hα	6006	0.9137	0.9882	0.9744	0.2220	0.0845	0.1231
HB	1425	0.9828	0.9995	0.9939	0.1952	0.0360	0.1165
HB2	3417	0.9401	0.9957	0.9816	0.2538	0.0697	0.1424
HB3	2213	0.9346	0.9952	0.9783	0.2689	0.0745	0.1566
HD2	1483	0.9951	0.9996	0.9963	0.2355	0.0557	0.2049
HD3	640	0.9658	0.9944	0.9849	0.2434	0.0988	0.1627
HE2	573	0.9956	0.9995	0.9965	0.1878	0.0607	0.1668
HE3	295	0.9934	0.9967	0.9959	0.1883	0.1884	0.1494
HG2	2234	0.9569	0.9581	0.9761	0.1927	0.1190	0.1445
HG3	1200	0.9107	0.9867	0.9580	0.2067	0.0812	0.1436

Table S10. Relative influence of various features for developing prediction models of SHIFTX+. (AA = amino acid type; SS = secondary structure; ASA=accessible surface area; SFE=salvation free energy; HA1/HA2/HN/OR dist=hydrogen bond length)

Feature	$^{13}C'$	$^{13}C\alpha$	$^{13}C\beta$	^{1}HN	$^{1}H\alpha$	^{15}N
AA_i	0.6	11.6	15.4	0.5	0.8	3.4
AA_{i+1}	2.3	1.0	0.3	0.3	0.7	0.3
AA_{i+2}	0.3	0.1	0.1	0.3	0.1	0.2
AA_{i-1}	0.4	0.1	0.1	0.4	0.2	2.9
AA_{i-2}	0.3	0.1	0.1	0.3	0.2	0.2
SS_i	8.1	0.1	0.1	0.1	0.6	0.0
SS_{i+1}	0.2	0.1	0.2	0.1	0.1	0.1
SS_{i-1}	0.1	0.0	0.0	0.1	0.1	0.1
φ_i	5.8	11.0	8.1	4.4	29.9	4.5
φ_{i+1}	3.6	1.1	0.6	0.9	1.3	0.9
φ_{i+2}	0.6	0.2	0.1	0.7	0.7	0.7
φ_{i-1}	0.6	0.4	0.3	2.1	1.0	2.1
φ_{i-2}	0.6	0.3	0.4	2.3	0.5	1.3
ψ_i	13.9	10.4	5.7	5.3	3.8	7.1
ψ_{i+1}	8.6	0.9	0.3	0.6	0.6	0.5
ψ_{i+2}	1.6	0.3	0.1	0.5	0.3	0.4
ψ_{i-1}	1.4	0.3	0.2	15.3	0.4	18.7
ψ_{i-2}	0.4	0.2	0.2	5.9	0.4	0.5
χ_i	4.1	2.6	1.3	0.8	1.3	5.9
χ_{i+1}	1.1	0.3	0.2	0.7	0.6	0.4
χ_{i+2}	0.8	0.2	0.2	0.6	0.3	0.5
χ_{i-1}	0.6	0.4	0.2	1.1	0.4	3.3
χ_{i-2}	0.6	0.2	0.2	0.7	0.3	0.4
$\chi 2_i$	3.1	2.2	1.4	0.5	0.4	1.6
$\chi 3_i$	0.1	0.1	0.1	0.2	0.1	0.1
ω_i	0.6	0.1	0.4	0.7	0.2	0.4
ω_{i+1}	0.2	0.1	0.1	0.6	0.2	0.3
ω_{i+2}	0.3	0.1	0.1	0.3	0.3	0.2
ω_{i-1}	0.3	0.3	0.9	0.4	0.7	0.4
ω_{i-2}	0.3	0.1	0.1	0.3	0.2	0.3
κ_i	2.5	0.3	0.2	3.1	0.4	0.4
θ_i	2.3	0.6	0.3	5.3	0.8	0.5

ASA	0.1	0.4	0.4	0.0	0.0	0.2
Residue ASA	4.2	0.3	0.2	1.2	0.6	0.5
Residue frac. ASA	0.6	0.1	0.0	0.6	0.2	0.1
Side chain ASA	0.6	0.2	0.1	0.5	0.3	0.4
Side chain frac. ASA	0.5	0.1	0.1	0.7	0.3	0.2
SFE	0.9	0.5	0.2	0.6	1.0	0.6
HA1 dist $_{i}$	0.0	0.0	0.0	0.0	0.0	0.0
HA1 dist $_{i+1}$	0.0	0.0	0.0	0.0	0.0	0.0
HA1 dist $_{i-1}$	0.0	0.0	0.0	0.0	0.0	0.0
HA2 dist $_{i}$	0.0	0.0	0.0	0.0	0.0	0.0
HA2 dist $_{i+1}$	0.0	0.0	0.0	0.0	0.0	0.0
HA2 dist $_{i-1}$	0.0	0.0	0.0	0.0	0.0	0.0
Hbond status $_{i}$	0.1	0.1	0.1	0.6	0.2	0.1
Hbond status $_{i+1}$	0.5	0.2	0.1	0.1	0.7	0.3
Hbond status $_{i-1}$	0.0	0.1	0.1	0.3	0.7	0.0
HN dist $_{i}$	0.4	0.1	0.1	0.3	0.2	0.2
HN dist $_{i+1}$	0.3	0.1	0.1	0.3	0.2	0.4
HN dist $_{i-1}$	0.3	0.1	0.1	0.5	0.3	0.3
OR dist $_{i}$	0.2	0.1	0.1	0.3	0.3	0.2
OR dist $_{i+1}$	0.2	0.1	0.1	0.3	0.2	0.2
OR dist $_{i-1}$	0.2	0.1	0.1	0.2	0.2	0.2
Hydrophobicity	0.5	0.2	0.1	0.6	0.3	0.4
Random coil	22.5	50.0	58.5	3.0	21.3	35.9
Ring current	0.0	0.5	0.9	11.5	11.2	0.6
Electric field effect	0.0	0.3	0.0	2.7	12.9	0.0
Hbond effect	0.0	0.0	0.0	18.4	0.3	0.0
pH	0.1	0.0	0.6	0.6	0.2	0.0
Temperature	0.0	0.0	0.0	0.5	0.1	0.0