Supporting Information

A Novel Approach to Protein Chemical Shifts Prediction from Sequence Using Protein Language Model

He Zhu, Lingyue Hu, Yu Yang*, and Zhong Chen*

Department of Electronic Science, Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, State Key Laboratory of Physical Chemistry of Solid Surfaces, Xiamen University, Xiamen, China

Corresponding authors

*E-mail: yuyang15@xmu.edu.cn, and chenz@xmu.edu.cn

Table of Contents

S1. Performances of PLM-CS and comparisons with SHIFTX2

Table 1. RMSE values of PLM-CS on each sample in the *SHIFTX test set*

Index	BMRB id	Cα	Cβ	C	Hα	H	N
1.	4032	1.3057	1.4206	0.7849	0.1502	0.2607	2.6560
2.	6338	1.3748	1.4334	1.0547	0.3719	0.4657	3.2506
3.	4834	1.2695	1.4089	1.2216	0.3576	0.5503	3.3278
4.	6597	1.0587	0.8918	0.9028	0.2585	0.3743	3.5523
5.	6709	0.8799	1.6989	0.8105	0.2635	0.2575	2.6148
6.	6344	1.4954	1.6251	1.3704	0.3859	0.6324	3.5007
7.	6457	0.3752	0.2626	0.7073	0.0623	0.0706	0.7952
8.	4053	0.6003	0.7738	0.6549	0.2185	0.1806	0.8836
9.	6292	1.2592	1.3889	1.2312	0.3407	0.4999	3.0068
10.	6198	1.0480	0.8234	0.9317	0.2432	0.3466	2.5922
11.	6114	1.4098	1.1206	1.2831	0.3612	0.6165	3.1255
12.	6635	1.0526	0.8744	0.7477	0.3678	0.3547	2.0694
13.	6223	1.7197	1.3686	1.5583	0.3808	0.7882	3.4581
14.	15757	1.5683	1.7077	1.3953	0.3632	0.6454	3.7687
15.	16173	1.4076	1.7190	1.4546	0.6115	0.6351	3.8685
16.	7322	0.8286	0.7277	0.6618	0.3035	0.3333	1.6191
17.	15517	0.9587	1.4983	0.9483	0.2785	0.2564	1.4793
18.	4879	1.4815	1.6144	1.0557	0.4040	0.5598	3.7331
19.	15249	0.8637	0.9944	1.0564	0.1967	0.3979	2.6809
20.	6252	0.9219	0.8839	0.8914	0.2106	0.3244	2.4441
21.	15243	1.5107	1.6008	1.2892	0.3541	0.5779	3.4433
22.	7114	0.5005	0.4227	0.2712	0.2575	0.1671	0.7966
23.	16007	1.3442	1.8034	1.2709	0.3705	0.5813	3.6644
24.	15089	1.3235	0.8517	1.1591	0.3004	0.5004	2.3813
25.	4113	1.1797	1.2621	1.2449	0.3141	0.5134	3.2885
26.	15281	1.2996	1.2232	1.0984	0.3343	0.5031	2.6794
27.	15530	1.1643	1.2736	1.0514	0.3313	0.5100	3.1613
28.	7055	1.2772	1.2324	1.3773	0.3807	0.5538	2.9512
29.	4300	0.2942	0.2577	\	0.0917	0.2596	1.6957
30.	6357	0.6188	0.9233	0.4519	\	0.0809	0.9822
31.	6197	0.9094	1.2132	1.1845	0.2628	0.5320	2.7558
32.	4047	1.0663	0.9164	1.1029	0.3323	0.3700	2.5155
33.	4149	1.3829	1.3194	1.0934	0.4391	0.6022	2.9378
34.	6032	0.4295	0.4314	0.5178	0.1689	0.1883	0.6204
35.	6604	0.5308	0.4601	0.5312	0.2118	0.1890	1.0445
36.	5358	0.5202	\	0.8019	0.1774	0.2825	0.8311
37.	4083	0.5101	0.9592	\	0.0963	0.1276	0.9350
38.	6321	2.0440	3.0083	\	0.3412	0.4699	2.1273
39.	6313	1.6678	0.9843	2.6702	0.3249	0.7277	3.1287
40.	6391	1.1861	1.1475	0.8630	0.3807	0.4150	3.1969
41.	6034	1.3056	1.7584	1.1943	\	0.5945	3.9036
42.	4037	1.1974	\	1.2400	0.3101	0.5215	3.2401

43.	5967	0.7298	0.9674	\	\	0.3780	2.8496
44.	6277	0.9421	0.7078	0.8704	0.2191	0.1453	0.8123
45.	6193	1.0606	1.1206	0.8433	0.3039	0.4045	3.1942
46.	6132	1.3702	1.5259	1.3417	0.3742	0.6539	3.2702
47.	4560	2.4198	1.7492	1.7437	0.4225	0.5554	3.7824
48.	2208	1.0810	\	1.0457	0.2971	0.2771	1.9400
49.	4831	1.0458	0.7913	0.9543	\	\	3.0135
50.	6250	0.3020	0.2057	\	0.2015	0.1299	0.6392
51.	5760	0.8380	1.1320	\	\	0.3716	2.6610
52.	6560	0.6739	0.6740	\	0.1862	0.2185	1.1911
53.	6024	0.7647	1.0275	0.8472	\	0.1426	1.0875
54.	6016	1.5703	1.8569	1.3460	\	0.6706	4.2575
55.	15560	1.1423	1.2591	1.1146	0.4732	0.1287	1.1254
56.	15232	0.8242	0.8986	\	0.2330	0.1574	0.9870
57.	7242	1.1986	1.1358	\	0.3251	0.4640	2.9164
58.	15501	1.2038	1.5051	1.0898	\	0.4205	2.8942
59.	15741	1.4701	1.2816	1.1714	\	0.5849	3.5445
60.	10139	0.6253	0.7937	0.6560	\	0.1837	0.7924
61.	10053	1.6347	1.2531	1.3234	\	0.5774	3.7872

Table 2. RMSE values of SHIFTX2 on each sample in the *solution-NMR test set*

Index	BMRB id	PDB id	Cα	Cβ	C	Hα	H	N
1.	11468	2rsc	1.1907	1.4341	1.2614	0.3044	0.5650	3.2039
2.	11469	2rsd	1.1646	0.8741	1.3111	0.2633	0.4732	3.0191
3.	11471	2rse_A	0.4310	0.5126	\	0.2941	0.3716	3.5539
4.	11471	2rse_B	1.7785	1.8613	\	0.2979	0.8812	6.2986
5.	11472	2rsf	1.0625	1.0541	0.9942	0.2693	0.5169	3.1863
6.	11473	2rsg	1.2321	1.3462	1.1791	0.4227	0.6148	3.2006
7.	11487	2rv7	0.8123	1.0624	0.8928	0.2267	0.4861	2.6361
8.	11488	2rv6	0.9557	0.9449	0.7686	0.1818	0.4483	3.0411
9.	11478	2ruy	1.3914	1.3874	1.1355	0.3102	0.5863	3.2430
10.	11491	2rsm	1.0836	1.0336	1.0658	0.2365	0.4530	3.1981
11.	11504	2rt5_A	0.9940	1.1209	1.0171	0.2835	0.5241	3.2038
12.	11506	2rsv	1.4319	1.7714	1.3433	\	0.6249	3.5822
13.	11508	2rsy_A	1.3058	1.3877	1.1150	0.3221	0.7069	2.9389
14.	11523	2rt3	1.0503	1.0823	1.3624	0.4279	0.5412	3.1117
15.	11525	2rt6	0.8255	0.8468	1.1078	0.2053	0.5665	2.4714
16.	11530	2rts	1.1803	2.2939	1.1125	0.3169	0.6159	2.6369
17.	11531	2rtt	1.0695	1.4709	1.2725	0.5530	0.6575	3.9207
18.	11534	2rtx	1.1776	1.1395	1.1186	0.2491	0.5432	3.4585
19.	16045	2kbi	2.3554	3.0080	2.4287	0.2389	0.5258	2.4007
20.	16050	2kbm_B	1.3144	2.5036	1.2769	0.2294	0.5220	2.9112
21.	16053	2kbt	1.3019	1.5851	1.1202	0.4422	0.5556	3.3835
22.	16061	2kc5	1.2131	1.7862	1.3345	0.4027	0.5116	2.9956

23.	16066	2kc9	1.0172	0.9129	0.8850	0.2397	0.4199	2.2966
24.	16068	2kca	\	\	\	0.3955	0.5926	\
25.	16082	2kcj	0.9898	1.2843	0.9761	0.3933	0.4671	3.0651
26.	16091	2kcp	1.0172	0.9897	\	0.2512	0.4107	3.0071
27.	16099	2kcx	1.0257	1.1903	1.0040	0.2301	0.5615	3.1678
28.	16098	2kcw	1.0697	1.0672	1.1476	0.3392	0.5403	2.5075
29.	16114	2kdi	2.7076	2.8489	\	0.3114	0.4115	1.9053
30.	16117	2kdm	1.0739	1.0744	0.8767	0.2664	0.5115	3.6039
31.	16116	2kdl	1.5057	1.0638	1.2030	0.3807	0.5507	4.1537
32.	17639	2lcz	1.4249	1.0889	1.2312	0.3654	0.4099	2.6352
33.	17643	2mwg_A	1.2322	1.2165	1.0444	\	0.5639	2.5735
34.	17667	2ll4_M	1.2489	1.3790	1.6623	0.3345	0.5718	3.0419
35.	17707	2lec_A	2.9319	3.0860	\	0.3428	0.7181	4.3310
36.	17723	2leq	1.1574	0.9618	0.9424	0.4138	0.6198	3.3226
37.	17711	2leh_A	1.3787	1.2198	1.0750	0.3336	0.6067	2.5988
38.	17752	2lfg	1.0676	1.3460	1.2605	0.3249	0.4911	3.1373
39.	17738	2lf4	1.2592	1.5016	1.3025	0.2775	0.4845	2.6794
40.	4023	1q2n	0.7144	0.9349	1.3507	0.0901	0.2160	1.0658
41.	4198	1br0	0.8898	0.6852	0.6401	0.1886	0.3198	1.5606
42.	4225	3grx	\	\	\	0.2328	0.5394	2.9925
43.	4264	1vrc_A	1.4750	3.2334	0.6177	1.1031	0.6891	2.4240
44.	4264	1vrc_D	4.2534	3.6885	3.8798	0.7591	0.5704	2.2885
45.	5467	1ri9	1.4945	1.4526	1.2817	0.4043	0.5703	3.1682
46.	5461	1lui	1.4356	1.5230	\	0.3717	0.4860	3.1477
47.	5480	1sy9_A	0.6397	0.6139	0.6779	0.1906	0.4298	1.2942
48.	5482	1mjd	1.1516	1.2293	\	0.2822	0.4498	2.8237
49.	5498	1k4u_S	1.3204	1.5583	1.3219	0.4173	0.6764	3.6385
50.	6111	1z2f	2.9887	2.6247	2.9077	0.3743	0.5589	3.7953
51.	6114	1oqa	1.2060	1.2250	0.6441	0.3461	0.6008	2.9283
52.	6161	1ust	0.8842	0.8252	\	0.2277	0.4130	3.0311
53.	6173	2lp6	1.0049	1.3163	0.9240	0.3534	0.4648	3.0465
54.	6201	1t6w	\	\	\	0.3185	0.4776	3.1508
55.	6211	1tdp	1.2838	1.1267	1.3295	0.2600	0.6166	2.5679
56.	6295	1tvc	1.5975	1.7922	3.7029	0.4795	0.8236	3.6244
57.	6800	2gg1	0.7037	1.1244	0.7892	0.2993	0.5793	2.3438
58.	6573	1yxr	1.0375	0.8072	\	0.2229	0.5588	2.2617
59.	6806	2b87_B	0.5405	0.6928	0.7620	0.1561	0.4426	1.5823
60.	6895	2err_A	1.6278	1.4972	\	0.3644	0.7258	4.0480
61.	6829	2czn	1.1794	1.2803	1.1819	0.3652	0.5769	3.6675
62.	6845	2aqc	0.9975	0.9860	0.9868	0.2089	0.3825	2.5220
63.	6921	2i96	1.9268	1.4741	1.2966	0.5427	0.8351	3.3937
64.	6934	2fek	1.0214	1.2536	1.1488	0.3013	0.5661	2.6449
65.	7053	2v31	1.4430	1.4086	1.1018	0.4025	0.7140	3.3897
66.	7061	2g35_A	1.0787	1.1407	\	0.3618	0.4689	3.0226
67.	7093	2gtv	1.4854	1.4850	\	0.2699	0.5665	3.1140

68.	7116	2glw	0.7503	1.2178	1.0191	0.3088	0.4979	2.1827
69.	7125	2m6z_A	1.2263	1.0951	\	0.2991	0.4588	2.5036
70.	7125	2m6z_B	1.2914	1.0290	\	0.3665	0.3940	1.1347
71.	7150	2h7d_A	0.7937	1.0257	1.2105	0.3541	0.4860	2.6650
72.	7158	2gqe	1.3729	1.1898	1.6195	0.1905	0.5364	4.3711
73.	7178	2gzp	\	3.0386	\	0.3523	0.5542	2.7378
74.	15058	2jms	\	\	\	0.4176	0.6576	\
75.	15592	2jy8	1.3152	1.1621	1.1830	0.2743	0.5409	2.5421
76.	15593	2lgr	1.2702	1.1788	\	0.2550	0.4763	3.2229

Table 3. RMSE values of PLM-CS on each entry of the *solution-NMR test set*

Index	BMRB id	Cα	Cβ	C	Hα	H	N
	11468	1.3389	1.0532	1.2436	0.3240	0.4172	2.7307
	11469	1.0773	0.9734	1.2504	0.2847	0.4137	4.1339
	11471	0.6573	0.2550	\	0.2955	0.1029	2.0747
	11471	0.6573	0.2550	\	0.2955	0.1029	2.0747
	11472	1.1241	0.9321	0.9576	0.3993	0.3878	2.7841
	11473	1.2522	0.9242	0.7818	0.4418	0.4214	2.7448
	11487	1.1210	0.7632	0.9131	0.2119	0.3264	2.8089
	11488	1.0728	0.7334	0.8732	0.2030	0.3641	2.8036
	11478	1.1478	2.0303	1.4091	0.2498	0.2677	2.4911
	11491	1.0534	0.8408	0.9591	0.2317	0.3369	2.2839
	11504	1.4279	1.3518	1.3071	0.3651	0.5800	3.7609
	11506	1.4456	1.4271	1.3088	\	0.6234	3.2703
	11508	1.3467	2.5815	1.0543	0.3217	0.4824	2.7164
	11523	0.8830	0.8511	1.2445	0.2990	0.3332	2.6880
	11525	1.4618	0.8751	1.2211	0.2314	0.5517	2.7697
	11530	0.8695	2.1892	0.8334	0.2977	0.4014	2.2691
	11531	1.5112	1.2774	1.2735	0.4918	0.6234	3.7177
	11534	1.0419	0.8920	0.9913	0.2301	0.4024	2.6402
	16045	2.2539	2.8423	2.4855	0.2637	0.5234	2.1882
	16050	1.5247	2.4723	1.3668	0.2279	0.1989	2.1607
	16053	0.9651	0.7000	0.7710	0.2477	0.2983	1.7135
	16061	1.4095	2.7011	1.2933	0.4093	0.4981	3.3431
	16066	1.3195	1.2389	1.1290	0.3695	0.5158	2.6527
	16068	\	\	\	0.3580	0.4571	\
	16082	1.0752	0.8164	0.8809	0.2849	0.3651	2.5936
	16091	1.1930	1.2851	\	0.3881	0.4245	2.5803
	16099	1.1601	0.9202	1.0419	0.2720	0.4108	3.0069
	16098	1.3652	1.3318	1.1998	0.3413	0.5042	3.0023
	16114	2.4703	2.7326	\	0.2462	0.2364	1.6141
	16117	2.3827	2.1672	2.3060	0.7435	0.8853	4.8162
	16116	0.8000	0.6305	0.9692	0.1829	0.2881	2.3894
	17639	1.0921	0.7836	1.2036	0.2525	0.2384	1.6244

17643	1.2998	1.4106	1.0802	\	0.4421	2.8824
17667	1.3579	2.4290	1.5166	0.4486	0.5175	3.2100
17707	2.5824	2.6094	\	0.2132	0.3652	2.1132
17723	1.2268	1.0121	1.0825	0.3538	0.6297	3.2822
17711	1.4576	1.9496	1.1741	0.3444	0.4573	2.5790
17752	0.9317	2.3343	1.1946	0.2781	0.4309	2.7841
17738	1.4734	1.3566	1.4804	0.2704	0.4270	2.4631
4023	0.8126	0.5977	1.8490	0.1322	0.1205	1.0431
4198	0.7532	0.5364	0.6897	0.1540	0.1714	1.4662
4225	\	\	\	0.1576	0.4717	2.0732
4264	0.5871	0.7164	1.5332	0.1534	0.1463	2.6580
4264	0.5871	0.7164	1.5332	0.1534	0.1463	2.6580
5467	1.5450	1.1439	\	0.3867	0.4970	2.8571
5461	1.0752	1.1336	\	0.2630	0.3625	2.4241
5480	0.7686	0.4703	0.6451	0.2072	0.1873	0.6940
5482	1.4372	1.2457	\	0.3393	0.4548	2.8433
5498	1.0698	0.7439	0.8016	0.2168	0.4134	2.0427
6111	3.3986	3.2776	2.9792	0.5887	0.5180	4.0613
6114	1.5573	1.1154	1.2526	0.3872	0.6165	3.0698
6161	1.0187	0.5231	\	0.2122	0.3214	2.3693
6173	1.0129	1.0856	0.9045	0.4099	0.3444	2.8273
6201	\	\	\	0.3233	0.4211	3.0233
6211	1.4079	0.9020	1.1661	0.2431	0.4159	2.7244
6295	1.3571	1.5408	3.6863	0.3697	0.5934	3.4263
6800	1.1784	1.1603	0.9769	0.3092	0.3626	2.6014
6573	1.0017	0.6372	\	0.1785	0.3145	1.6326
6806	1.0694	0.4099	0.4522	0.1715	0.2101	1.1991
6895	1.1237	1.2455	\	0.3195	0.4875	2.5297
6829	1.2499	1.3052	1.1489	0.3886	0.5523	3.0889
6845	1.2353	1.4696	0.9933	0.2693	0.3065	2.5193
6921	2.0919	6.1080	1.3052	0.7477	0.9282	3.2707
6934	1.2597	1.0245	1.1359	0.3275	0.4914	2.6425
7053	1.4220	1.2206	1.0922	0.3609	0.6897	3.3526
7061	1.6388	1.1013	\	0.3143	0.4290	3.5434
7093	1.5189	1.4793	\	0.2779	0.3911	2.6374
7116	1.0190	1.1941	1.0586	0.2851	0.3971	2.7514
7125	1.3592	1.0811	\	0.2639	0.2857	2.2189
7125	1.3592	1.0811	\	0.2639	0.2857	2.2189
7150	1.0232	1.1782	1.1757	0.3412	0.4413	3.1329
7158	1.3790	1.6463	1.6566	0.2178	0.3689	4.0996
7178	4.1007	3.2263	\	0.3439	0.2986	2.0415
15058	\	\	\	0.5102	0.6749	\
15592	1.0799	0.6656	1.0402	0.2439	0.2525	2.8145
15593	1.5466	1.4782	\	0.3886	0.4942	3.3056

Table 4. RMSE values of SHIFTX2 with AlphaFold-predicted structures on the *solution-NMR test set.* Removed samples: BMRB id: 16068, Cα, Cβ, C, N; BMRB id: 15593, C; BMRB id: 6812, Cα.

Index	BMRB id	PDB id	Cα	Cβ	C	Hα	H	N
1.	11468	2rsc	1.4933	1.0108	1.0108	0.4575	0.4575	2.4786
2.	11469	2rsd	1.1605	1.3270	1.3270	0.5148	0.5148	3.6407
3.	11471	2rse_A	0.3867	0.5579	\	0.2790	0.2771	3.3812
4.	11471	2rse_B	2.1454	1.6454	\	0.2942	0.5231	5.2345
5.	11472	2rsf	0.9711	0.8326	0.9392	0.2641	0.3744	2.6816
6.	11473	2rsg	1.1144	1.0725	1.1116	0.2973	0.4515	2.7989
7.	11487	2rv7	1.1296	1.1285	1.0962	0.2374	0.5593	3.1272
8.	11488	2rv6	1.0887	1.1340	1.0388	0.2443	0.5466	3.4669
9.	11478	2ruy	2.0417	2.4496	1.7929	0.2737	0.4917	3.4304
10.	11491	2rsm	0.9612	0.9411	0.7810	0.2212	0.3759	2.5952
11.	11504	2rt5_A	1.3282	0.8142	0.9055	0.1971	0.4269	3.4066
12.	11506	2rsv	1.0579	1.2506	1.1894	\	0.5093	2.5483
13.	11508	2rsy_A	0.9683	2.4620	0.9895	0.2575	0.5365	2.3243
14.	11523	2rt3	1.0282	0.9273	1.3500	0.2577	0.5042	2.9809
15.	11525	2rt6	0.8441	0.6827	1.0558	0.2202	0.5112	2.1685
16.	11530	2rts	1.1100	2.3183	0.9855	0.3219	0.4984	2.8192
17.	11531	2rtt	0.8046	1.3185	1.0609	0.2809	0.3511	2.8476
18.	11534	2rtx	0.9184	0.8605	0.7622	0.1853	0.4372	2.8828
19.	16045	2kbi	2.9141	2.9661	2.7277	0.2208	0.5689	2.2190
20.	16050	2kbm_B	3.9673	2.5351	2.9922	0.4809	0.6671	5.2386
21.	16053	2kbt	3.3014	3.0430	2.5412	0.8294	0.9711	4.3343
22.	16061	2kc5	0.9465	1.5329	1.2242	0.3035	0.4526	2.4685
23.	16066	2kc9	1.0179	0.7556	0.9144	0.2000	0.3634	2.2576
24.	16068	2kca	\	\	\	0.2759	0.4900	\
25.	16082	2kcj	1.1299	1.0213	1.0924	0.2447	0.3902	2.2872
26.	16091	2kcp	1.0281	1.0240	\	0.2768	0.3659	2.3733
27.	16099	2kcx	0.9291	1.1847	1.0523	0.2017	0.3730	2.5651
28.	16098	2kcw	0.8440	0.9902	1.0375	0.2931	0.4686	1.9853
29.	16114	2kdi	2.8443	2.6353	\	0.2198	0.4191	2.2577
30.	16117	2kdm	1.8705	1.4760	1.7287	0.2313	0.4150	2.0023
31.	16116	2kdl	1.4661	1.3844	1.2749	0.2415	0.3983	3.3162
32.	17639	2lcz	1.4089	2.7957	1.9500	0.2425	0.3987	2.6720
33.	17643	2mwg_A	1.0104	1.3799	0.9376	\	0.4009	2.3605
34.	17667	2ll4_M	0.9268	1.2604	1.5840	0.2555	0.4591	2.7121
35.	17707	2lec_A	2.9189	2.4957	\	0.1931	0.4610	2.5563
36.	17723	2leq	0.8169	1.4810	1.3561	0.2501	0.4621	2.7393
37.	17711	2leh_A	0.8598	0.9604	0.8672	0.2031	0.4001	2.1008
38.	17752	2lfg	0.9242	1.1280	1.3404	0.2429	0.4465	3.1887
39.	17738	2lf4	1.1887	1.3746	1.3611	0.1935	0.3779	2.1101
40.	4023	1q2n	1.2635	0.8114	1.7547	0.1990	0.3689	2.4340
41.	4198	1br0	0.6280	0.5168	0.4335	0.0820	0.1588	1.4240

42.	4225	3grx	\	\	\	0.1799	0.4392	2.3282
43.	4264	1vrc_A	2.9522	2.1850	2.4936	0.6711	0.8505	4.8277
44.	4264	1vrc_D	2.9522	2.1850	2.4936	0.6711	0.8505	4.8277
45.	5467	1ri9	1.3298	1.2961	1.0614	0.2285	0.5063	2.8094
46.	5461	1lui	0.9995	1.1258	\	0.2371	0.4863	3.0128
47.	5480	1sy9_A	0.6354	0.5836	0.6778	0.1293	0.2877	0.8568
48.	5482	1mjd	1.1655	1.1283	\	0.2450	0.4286	2.3462
49.	5498	1k4u_S	1.0929	1.0937	1.0153	0.2858	0.5387	2.5561
50.	6111	1z2f	2.8759	2.4831	2.7611	0.3337	0.4551	2.8534
51.	6114	1oqa	1.0324	0.8760	0.8992	0.1881	0.4533	2.8104
52.	6161	1ust	0.7516	0.7630	\	0.1510	0.3871	2.2949
53.	6173	2lp6	0.8392	1.1468	0.8799	0.3040	0.3455	2.5109
54.	6201	1t6w	\	\	\	0.2633	0.3542	2.2710
55.	6211	1tdp	0.8265	0.7735	0.8994	0.2747	0.3804	12.3636
56.	6295	1tvc	1.0518	1.4330	3.8413	0.2696	0.6049	8.6796
57.	6800	2gg1	0.9278	1.3600	1.3147	0.2289	0.4020	1.9565
58.	6573	1yxr	1.0037	0.7422	\	0.1762	0.3405	2.0696
59.	6806	2b87_B	1.0108	0.9457	1.3056	0.2262	0.4008	2.2346
60.	6895	2err_A	1.2748	1.2017	\	0.2540	0.5637	2.4656
61.	6829	2czn	0.8923	1.0660	1.1029	0.3184	0.4725	2.6384
62.	6845	2aqc	1.4977	1.1761	1.4444	0.2149	0.5363	2.7955
63.	6921	2i96	1.6079	1.5117	1.1342	0.5325	0.8265	2.5271
64.	6934	2fek	0.9129	1.0456	1.1143	0.2577	0.5136	2.4358
65.	7053	2v31	1.4564	1.1416	1.0282	0.2572	0.6719	3.0986
66.	7061	2g35_A	1.1749	1.1639	\	0.2617	0.3363	2.2955
67.	7093	2gtv	1.0657	1.2166	\	0.1920	0.4903	2.2618
68.	7116	2glw	0.7474	1.0706	1.0164	0.2293	0.3887	1.9672
69.	7125	2m6z_A	0.8558	1.0245	\	0.2538	0.3676	1.8663
70.	7125	2m6z_B	0.8558	1.0245	\	0.2538	0.3676	1.8663
71.	7150	2h7d_A	1.3894	1.4463	1.3650	0.2283	0.3351	2.5716
72.	7158	2gqe	1.8166	4.2017	1.9381	0.3245	0.4930	4.2016
73.	7178	2gzp	4.2385	3.0896	\	0.2060	0.4454	2.1351
74.	15058	2jms	\	\	\	0.3816	0.5624	\
75.	15592	2jy8	1.0680	0.7082	1.5343	0.2503	0.4153	2.4780
76.	15593	2lgr	1.0310	1.1665	\	0.2362	0.3980	2.4686

Table 5. The average NMR chemical shift for each atom type of each amino acid

Amino acid type	Cα	Cβ	C	Hα	H	N
A	53.33	19.26	177.77	4.28	8.20	122.68
R	56.94	30.95	176.45	4,31	8.26	120.24
N	53.70	38.92	175.35	4.68	8.35	118.46
D	54.86	41.10	176.49	4.60	8.31	120.19
C	58.05	33.60	174.76	4.69	8.42	119.36
Q	56.75	29.41	176.37	4.28	8.22	119.34
E	57.52	30.24	176.98	4.27	8.35	120.14
G	45.57	\	173.95	3.90	8.33	109.13

H	56.65	30.59	175.28	4.62	8.27	119.13
I	61.69	38.90	175.85	4.23	8.30	121.12
L	55.74	42.56	177.03	4.35	8.24	121.45
K	57.11	33.04	176.71	4.28	8.20	120.52
M	56.27	33.30	176.24	4.42	8.27	119.57
F	58.27	40.23	175.49	4.65	8.40	120.10
P	63.49	32.06	176.74	4.40	\	129.88
S	58.82	63.98	174.67	4.51	8.30	115.86
T	62.33	69.85	174.58	4.49	8.27	115.06
W	57.85	30.37	176.22	4.70	8.30	121.31
Y	58.25	39.70	175.39	4.65	8.34	120.16
V	62.63	32.97	175.70	4.20	8.31	120.80

S2. Examples of using PLM-CS for validation of the chemical shift data

The proposed CS prediction method can be used to validate and screen chemical shift assignment results. By inputting a protein sequence, PLM-CS predicts the chemical shift values for each residue's backbone atoms. If these predicted values significantly differ from the given list of chemical shift assignment results, it may indicate potential misassignments. An example is shown in Figure S1a (BMRB id: 16068), where chemical shifts of some Cα atoms were assigned to ~170 ppm during the experiments while the predicted values should be around 62.5 ppm, indicating potential misassignments.

Another example is shown in Figure S1b, where the correlation between predicted values and experimental assignments is lower than that of the normal case shown in Figure S1c. We also present SHIFTX2 results in Figure S1d. Both our method and SHIFTX2 perform poorly for this protein, showing relatively large discrepancies between the predicted values and the assignments. This poor performance may be attributed to the nature of the protein, the hypothetical protein TA0938 from Thermoplasma acidophilum, which consists of flexible loops, unstructured coils, and mixed secondary structures. The inherent flexibility, lack of regularity, and complex local environments in this protein likely contribute to the difficulty in accurately predicting its chemical shifts.

While the proposed chemical shift prediction method demonstrates robust performance for many proteins, the challenges presented by proteins with complex structural features, such as flexible loops, unstructured coils, and mixed secondary structures, reveal areas for further refinement. These cases highlight the intricacies involved in accurately predicting chemical shifts under more demanding conditions. Nonetheless, the ability of our method to handle a broad range of proteins underscores its potential as a significant contribution to the field, laying a strong foundation for future advancements that can address these specific challenges.

These cases highlight that when using chemical shift prediction methods, it is important to recognize that discrepancies between assigned or expected values and predicted chemical shifts can arise for several reasons. First, such discrepancies might indicate a misalignment in the experimental data. Second, they may highlight the presence of unstructured regions within the protein, such as flexible loops or coils. Third, the local environment around certain residues might be unusually complex or dynamic, contributing to deviations in the predicted values. In practice, these potential sources of discrepancy can serve as valuable diagnostic tools, guiding further investigation into the protein's structure and dynamics. By understanding and accounting for these factors, users can more effectively interpret the results, leveraging the method's strengths while also being mindful of its limitations.

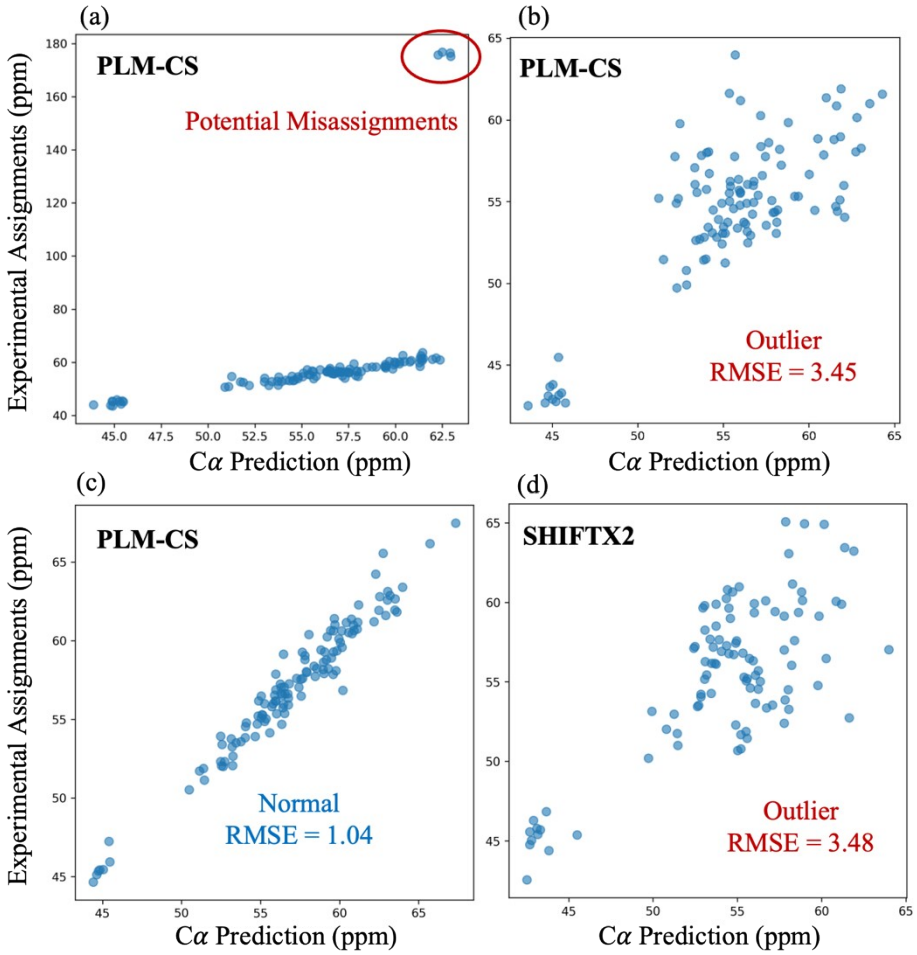

Figure S1. Three predictions of Cα values for two examples that were excluded from the test set due to the outliers. Cα, (a) Cα validation for an abnormal case (BMRB id: 16068); (b) Cα validation for an abnormal case (BMRB id: 6812); (c) Cα validation for a normal case (BMRB id: 11534); (d) Cα validation for an abnormal case (BMRB id: 6812, PDB id: 2FQH) using SHIFTX2.

S3. Examples of using PLM-CS for peak assignments

Figure S2 demonstrates three peak assignment examples. Figures S2a, c, and e show the PLM-CS predicted Cα-CO spectra, while Figures S2b, d, and f display the corresponding experimental peak distributions. To assist in experimental peak assignments, we select a specific peak (e.g., the k-th peak $S^{(k)} = \left(C_{\alpha}^{(k)}, C_{O}^{(k)}\right)$) in the experimental data (highlighted by yellow in Figures S2b, d, and f). We then calculate the probability that a peak in the predicted spectrum ($\tilde{S} = \left(\tilde{C}_{\alpha}, \tilde{C}_{O}\right)$) corresponds to $S^{(k)}$ as follows:

$$P\left(\tilde{S} = S^{(k)}\right) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left(- \frac{\left(\tilde{C}_{\alpha} - C_{\alpha}^{(k)}\right)^2}{\sigma_1^2} - \frac{\left(\tilde{C}_{O} - C_{O}^{(k)}\right)^2}{\sigma_2^2} \right)$$

where σ_1 and σ_2 represent the standard deviations of the predicted values $\tilde{C}_{\alpha}$ and $\tilde{C}_{O}$, estimated via statistical analysis of the test dataset. Figures S2a, c, and e use varying colors to indicate the likelihood of assignment for each peak, highlighting the residues most likely to match the selected peak.

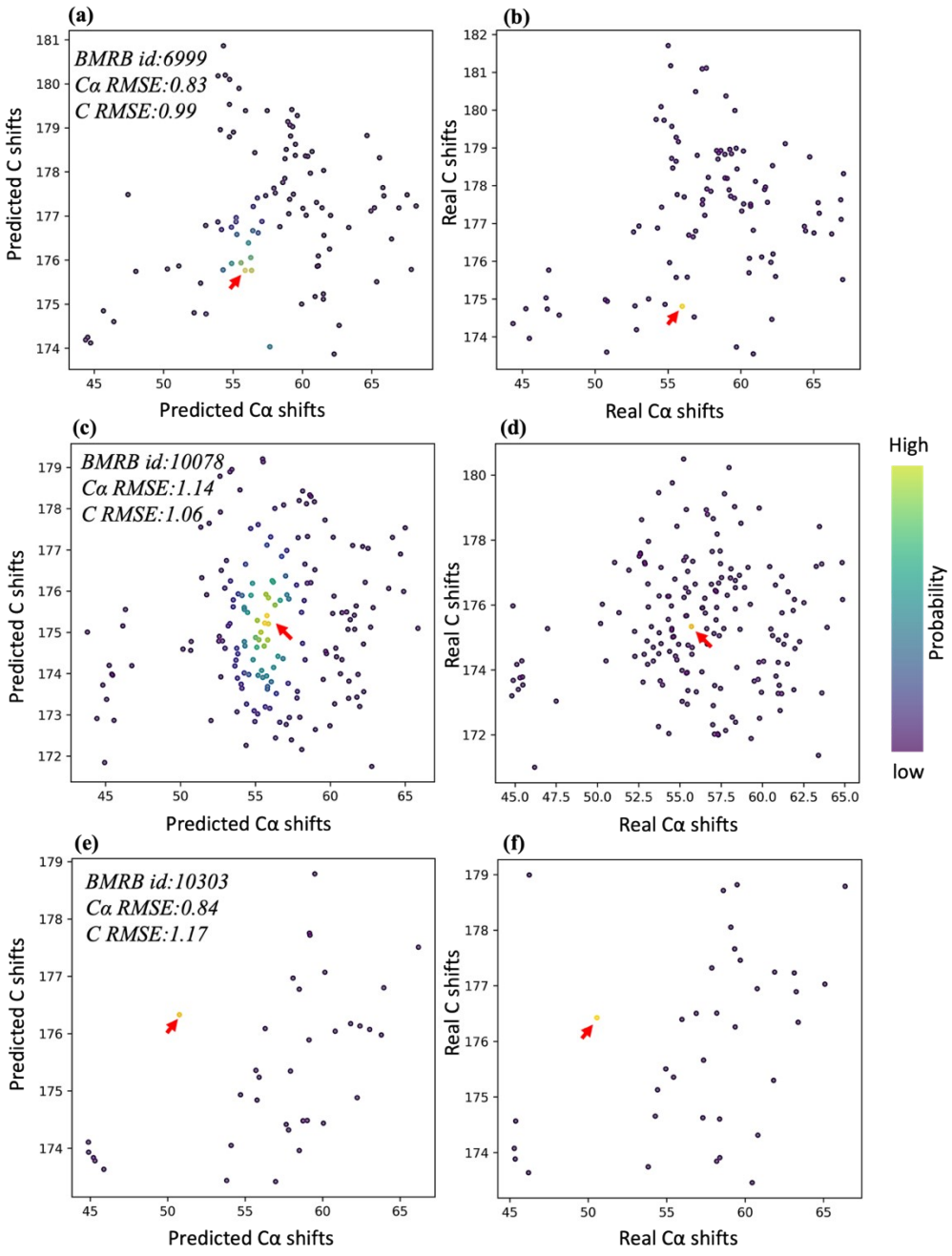

Figure S2. Three examples of PLMCS predicting chemical shifts. (a) and (b) are the predicted 2D spectra and the experimental chemical shifts for a protein with a BMRB ID 6999, respectively. The high bright spot indicated by the arrow in (b) represents a selected chemical shift pair. In (a), the color of each point indicates the probability that they belong to the same residue as the point specified in (b), and the point pointed by the arrow is the one that really meets the condition. (c) and (d) BMRB id 10078. (e) and (f) BMRB id 10303.

As shown in Figure S2, PLM-CS helps narrow down the amino acids needing attention and helps researchers in determining chemical shift assignments. In certain cases, it enables direct identification of specific chemical shift pairs.

However, it should be noted that this peak assignment approach based on predicted chemical shifts may address only a limited number of peaks in experimental data due to relatively high uncertainties in predictions and spectral crowding. In future work, we aim to develop a more comprehensive approach to integrate PLM-CS prediction results with the processing and interpretation of experimental spectra.

S4. Detailed structures of the Transformer predictor

The Transformer predictor proposed in this work features two Transformer encoders. Before entering the encoder, the data is projected from 1280 to 512 dimensions via a linear layer. The multi-head attention module in the Transformer encoder is shown in Figure S3a, where 8 attention heads and a 512-dimensional attention matrix are employed. The Feed-Forward module has the structure shown in Figure S3b, which consists of two linear layers with a Gaussian Error Linear Unit (GELU) as the activation function.

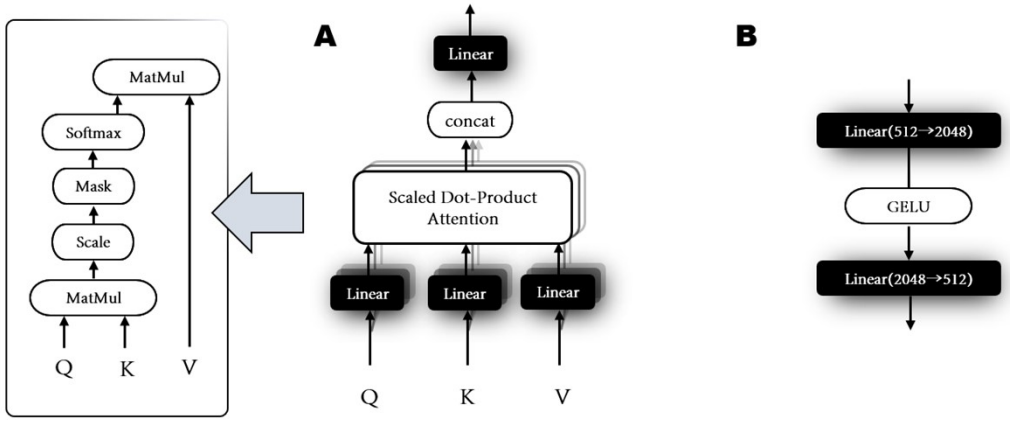

Figure S3. (a)The architecture of multi-head attention mechanism. (b) The architecture of Feed-Forward module.