# Supplementary Data for Accurate Prediction of Chemical Shifts for Aqueous Protein Structure on "Real World" Data

Jie Li[1,2], Kochise C. Bennett[1,2], Yuchen Liu[1,2], Martin V. Martin[3], Teresa Head-Gordon[1-4*]

*[1]Pitzer Center for Theoretical Chemistry, [2]Department of Chemistry, [3]Department of Bioengineering, [4]Department of Chemical and Biomolecular Engineering*
*University of California, Berkeley CA 94720*

## Exclusion of erroneous chemical shifts assignments in the test dataset

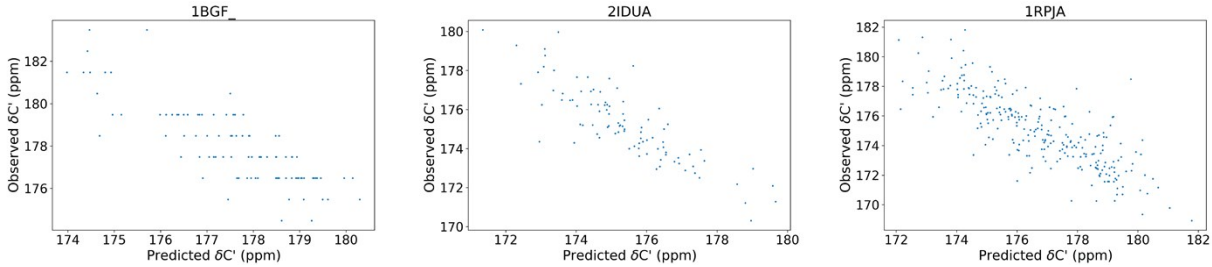
**Table S1.** Removed chemical shifts that are significantly offset from random coil average

| PDBID | RESID | RESNAME | Recorded CS (ppm) [atom] |
|---|---|---|---|
| 1MI4A | 153 | LEU | 0.09 [H] |
| 1DSBA | 30 | CYS | 331.45 [CB] |
| 2GYKE | 62 | ASP | 89.55 [CB] |
| 1GOA_ | 119 | TRP | 26.57 [N] |
| 1OVHA | 1 | MET | 38.5 [N] |
| 1D03A | 1 | ALA | 40.47 [N] |
| 2IGD_ | 1 | MET | 39.34 [N] |
| 1VB0A | 1 | LEU | 39.84 [N] |
| 1QOGA | 1 | ALA | 39.52 [N] |
| 1DSBA | 51 | PRO | 152.99 [N] |

**Table S2.** Excluded cysteine Cβ chemical shifts

| PDBID | RESID | Oxidation state in crystal structure | Recorded Cβ CS (ppm) |
|---|---|---|---|
| 2GOOB | 40 | reduced | 48.59 |
| 2PF5A | 47 | reduced | 43.71 |
| 1DSBA | 33 | oxidized | 33.6 |
| 1VKBA | 35 | oxidized | 31.16 |
| 1VKBA | 62 | oxidized | 26.76 |
| 2AFGB | 117 | reduced | 39.46 |
| 1LJUA | 82 | oxidized | 30.66 |
| 1LJUA | 89 | oxidized | 27.12 |

**Figure S1.** *Scatter plot for the C' of the three proteins whose C' shifts are removed.* Plotted are the observed C' chemical shifts versus predicted C' chemical shifts with UCBShift



## Compilation of test dataset

**Table S3.** *A complete list of the 200 testing proteins*. Provided are the PDB identifier, BMRB identifier, the X-ray resolution (RES.), highest sequence similarity to any example in the training dataset (SIM.), total number of residues and the number of residues and atom types with experimental chemical shifts.

Rows with green background are data with low sequence homology (<30%).

| | | | | Number of residues | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **PDB** | **BMRB** | **RES. (Å)** | **SIM.** | **Total** | **H** | **Hα** | **C'** | **Cα** | **Cβ** | **N** |
| 1MI4A | 4854 | 1.7 | 0.998 | 427 | 209 | 198 | 236 | 229 | 199 | 217 |
| 2GP0A | 15680 | 2.05 | 0.997 | 288 | 240 | 0 | 149 | 188 | 156 | 151 |
| 2H9HA | 4836 | 1.39 | 0.995 | 212 | 201 | 170 | 202 | 212 | 186 | 201 |
| 1T85A | 17415 | 1.8 | 0.995 | 406 | 305 | 0 | 0 | 299 | 196 | 233 |
| 2UYZA | 4132 | 1.4 | 0.994 | 156 | 137 | 142 | 0 | 141 | 107 | 116 |
| 1GOA_ | 4012 | 1.9 | 0.994 | 155 | 134 | 110 | 0 | 119 | 0 | 118 |
| 1SNO_ | 1878 | 1.7 | 0.993 | 136 | 0 | 0 | 0 | 0 | 0 | 109 |
| 5NUC_ | 5536 | 2.1 | 0.993 | 134 | 97 | 93 | 100 | 101 | 93 | 96 |
| 2EYOA | 16585 | 1.7 | 0.993 | 135 | 118 | 117 | 125 | 126 | 117 | 118 |
| 1NBPA | 6621 | 2.2 | 0.992 | 121 | 114 | 116 | 0 | 116 | 111 | 114 |
| 1LJUA | 4944 | 1.4 | 0.992 | 130 | 118 | 105 | 109 | 125 | 116 | 118 |
| 1IP2A | 5125 | 1.8 | 0.992 | 130 | 122 | 0 | 0 | 0 | 0 | 118 |
| 1AR0A | 5888 | 2.3 | 0.992 | 125 | 114 | 0 | 0 | 112 | 100 | 108 |
| 7TIMB | 16566 | 1.9 | 0.992 | 247 | 0 | 0 | 16 | 63 | 69 | 2 |
| 1OMSC | 15813 | 2.3 | 0.992 | 114 | 107 | 104 | 112 | 114 | 105 | 107 |
| 1BRJA | 975 | 2 | 0.991 | 107 | 101 | 94 | 0 | 0 | 0 | 0 |
| 2H76A | 62 | 2.25 | 0.991 | 108 | 88 | 79 | 0 | 0 | 0 | 0 |
| 2HSHA | 256, 257 | 1.35 | 0.990 | 105 | 59 | 60 | 0 | 0 | 0 | 0 |
| 1RN1B | 1658 | 1.84 | 0.990 | 103 | 0 | 0 | 0 | 0 | 0 | 96 |
| 2OCTA | 15548 | 1.4 | 0.990 | 97 | 84 | 0 | 89 | 88 | 83 | 84 |
| 1QOGA | 447 | 1.8 | 0.990 | 98 | 76 | 79 | 45 | 84 | 80 | 79 |
| 2GYKE | 4115, 4116 | 1.6 | 0.988 | 83 | 77 | 75 | 0 | 80 | 74 | 77 |
| 1POH_ | 29 | 2 | 0.988 | 85 | 82 | 79 | 0 | 0 | 0 | 0 |
| 1SPQB | 15066 | 2.16 | 0.988 | 239 | 185 | 0 | 0 | 183 | 146 | 147 |
| 1TPH2 | 15064 | 1.8 | 0.988 | 245 | 224 | 0 | 223 | 231 | 174 | 204 |
| 1J3FA | 4568 | 1.45 | 0.987 | 152 | 144 | 121 | 143 | 146 | 124 | 144 |
| 109M_ | 5158 | 1.83 | 0.987 | 154 | 129 | 0 | 131 | 142 | 0 | 129 |
| 1F2MA | 1875, 495 | 2 | 0.987 | 136 | 0 | 108 | 0 | 76 | 1 | 0 |
| 2IN8A | 16634 | 1.7 | 0.986 | 139 | 129 | 0 | 0 | 0 | 0 | 122 |
| 2FI4I | 4877 | 1.58 | 0.983 | 58 | 46 | 46 | 0 | 0 | 0 | 0 |
| 1P2OD | 45 | 2 | 0.983 | 58 | 47 | 44 | 0 | 0 | 0 | 0 |
| 2OW9B | 4679 | 1.74 | 0.982 | 166 | 137 | 140 | 145 | 154 | 139 | 137 |
| 1BNEA | 16169, | 2.1 | 0.982 | 107 | 84 | 55 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 16170 | | | | | | | | | |
| 1TXXA | 1812, 1813 | 2.2 | 0.981 | 108 | 0 | 0 | 0 | 0 | 0 | 95 |
| 2SGDI | 1375 | 1.8 | 0.980 | 51 | 0 | 0 | 0 | 43 | 15 | 0 |
| 1MOLB | 4633 | 1.7 | 0.979 | 94 | 77 | 73 | 0 | 0 | 0 | 0 |
| 1IV9A | 4222 | 1.9 | 0.979 | 96 | 84 | 80 | 0 | 0 | 0 | 78 |
| 1YJFA | 5514 | 1.35 | 0.979 | 225 | 174 | 96 | 174 | 198 | 136 | 174 |
| 2HWNB | 4473 | 1.6 | 0.978 | 45 | 39 | 41 | 0 | 43 | 41 | 39 |
| 2HZIB | 15488 | 1.7 | 0.975 | 264 | 244 | 0 | 36 | 41 | 0 | 37 |
| 1QKRA | 15653 | 1.8 | 0.973 | 172 | 146 | 0 | 139 | 146 | 126 | 130 |
| 2BC5C | 1672 | 2.25 | 0.972 | 106 | 91 | 92 | 0 | 0 | 0 | 91 |
| 1XRKB | 4785, 4786 | 1.5 | 0.968 | 121 | 104 | 104 | 103 | 110 | 104 | 104 |
| 1MYWA | 15826 | 2.2 | 0.962 | 228 | 192 | 123 | 184 | 189 | 139 | 175 |
| 1DS3I | 1374 | 1.65 | 0.961 | 50 | 39 | 38 | 0 | 0 | 0 | 0 |
| 1LW6I | 4974 | 1.5 | 0.953 | 63 | 57 | 55 | 56 | 57 | 54 | 57 |
| 1U06A | 7305, 7306 | 1.49 | 0.952 | 55 | 52 | 51 | 0 | 0 | 0 | 0 |
| 1SKOB | 6181 | 2 | 0.946 | 116 | 108 | 104 | 0 | 114 | 106 | 108 |
| 1P7JA | 7386 | 2.1 | 0.932 | 53 | 36 | 36 | 30 | 31 | 29 | 30 |
| 1AJ6_ | 5218 | 2.3 | 0.932 | 194 | 177 | 143 | 169 | 161 | 106 | 160 |
| 2H61H | 4105, 5895 | 1.9 | 0.924 | 90 | 80 | 79 | 78 | 83 | 78 | 80 |
| 1TCF_ | 1553 | 1.9 | 0.899 | 156 | 57 | 53 | 0 | 0 | 0 | 0 |
| 1KTZB | 5953, 5954, 4779 | 2.15 | 0.893 | 106 | 90 | 100 | 100 | 101 | 100 | 90 |
| 2B59A | 5267 | 2.11 | 0.872 | 166 | 157 | 142 | 132 | 153 | 121 | 138 |
| 1B68A | 6524 | 2 | 0.859 | 138 | 134 | 0 | 0 | 133 | 120 | 132 |
| 1BD9B | 17382 | 2.05 | 0.834 | 185 | 170 | 169 | 181 | 184 | 169 | 170 |
| 1N3ZA | 4980 | 1.65 | 0.815 | 117 | 108 | 108 | 0 | 0 | 0 | 101 |
| 1AG6_ | 79 | 1.6 | 0.788 | 99 | 86 | 83 | 0 | 0 | 0 | 0 |
| 1IKOP | 7220 | 1.92 | 0.764 | 141 | 127 | 120 | 127 | 133 | 122 | 127 |
| 2BIUX | 6141, 7310 | 1.71 | 0.745 | 164 | 139 | 123 | 0 | 146 | 127 | 139 |
| 1OFFA | 16024 | 1.8 | 0.735 | 95 | 91 | 88 | 0 | 0 | 0 | 91 |
| 1MR3F | 15626 | 1.6 | 0.718 | 177 | 145 | 0 | 0 | 135 | 100 | 123 |
| 1Z7XX | 4370 | 1.95 | 0.674 | 126 | 119 | 121 | 0 | 0 | 0 | 111 |
| 1GWYA | 16362, 16630 | 1.71 | 0.674 | 175 | 169 | 154 | 166 | 175 | 154 | 169 |
| 7RXN_ | 15374 | 1.5 | 0.673 | 52 | 44 | 43 | 0 | 0 | 0 | 44 |
| 1YCQA | 6248 | 2.3 | 0.645 | 88 | 75 | 72 | 0 | 78 | 66 | 72 |
| 1JHFB | 6373 | 1.8 | 0.644 | 111 | 96 | 0 | 89 | 101 | 83 | 90 |
| 1RDG_ | 5163 | 1.4 | 0.635 | 52 | 45 | 42 | 0 | 0 | 0 | 0 |
| 1PCS_ | 5475 | 2.15 | 0.580 | 98 | 84 | 80 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2AFGB | 15783, 16493, 16494, 16502, 6875 | 2 | 0.543 | 129 | 123 | 117 | 123 | 129 | 104 | 122 |
| 1WZVA | 16321 | 2.1 | 0.535 | 150 | 129 | 50 | 125 | 141 | 136 | 129 |
| 2GCNA | 16668 | 1.85 | 0.517 | 177 | 127 | 91 | 138 | 138 | 127 | 127 |
| 1L0SB | 5573 | 2.3 | 0.500 | 87 | 76 | 67 | 81 | 79 | 72 | 76 |
| 1OKHB | 4587 | 1.75 | 0.478 | 46 | 41 | 42 | 0 | 0 | 0 | 0 |
| 1OBOA | 5011 | 1.2 | 0.477 | 169 | 145 | 130 | 0 | 0 | 0 | 136 |
| 1RCF_ | 16593 | 1.4 | 0.472 | 169 | 98 | 91 | 0 | 2 | 0 | 43 |
| 1FDQA | 5320, 16042, 16046, 16047, 16048, 16049 | 2.1 | 0.466 | 131 | 128 | 121 | 0 | 130 | 121 | 128 |
| 1FU0A | 4774 | 1.9 | 0.448 | 86 | 84 | 77 | 0 | 0 | 0 | 0 |
| 1EK8A | 5190 | 2.3 | 0.438 | 185 | 175 | 0 | 181 | 181 | 0 | 167 |
| 2GGMB | 5503, 6687 | 2.35 | 0.436 | 137 | 130 | 121 | 0 | 0 | 0 | 62 |
| 1HFZC | 4811, 4332 | 2.3 | 0.435 | 121 | 111 | 99 | 0 | 0 | 0 | 110 |
| 1D03A | 1673 | 1.85 | 0.424 | 169 | 160 | 142 | 0 | 0 | 0 | 159 |
| 2HPWA | 16600 | 1.55 | 0.417 | 228 | 206 | 111 | 159 | 203 | 168 | 185 |
| 1YW5A | 16690 | 1.6 | 0.412 | 177 | 165 | 0 | 170 | 170 | 158 | 165 |
| 1GPR_ | 1663 | 1.9 | 0.407 | 158 | 146 | 133 | 0 | 0 | 0 | 134 |
| 2HPR_ | 932 | 2 | 0.402 | 86 | 79 | 74 | 0 | 0 | 0 | 0 |
| 5HPGA | 16466 | 1.66 | 0.393 | 84 | 75 | 73 | 0 | 0 | 0 | 75 |
| 1WYWB | 6304 | 2.1 | 0.381 | 79 | 77 | 0 | 0 | 0 | 0 | 71 |
| 1DSBA | 16327 | 2 | 0.381 | 188 | 0 | 0 | 146 | 167 | 124 | 151 |
| 1KDJ_ | 7370 | 1.7 | 0.373 | 102 | 96 | 0 | 0 | 0 | 0 | 90 |
| 1VB0A | 16060 | 0.92 | 0.348 | 61 | 0 | 0 | 53 | 55 | 50 | 51 |
| 1ZDNA | 17437 | 1.93 | 0.348 | 151 | 132 | 0 | 0 | 122 | 104 | 121 |
| 2AAOB | 5324 | 2 | 0.343 | 131 | 118 | 115 | 0 | 114 | 104 | 111 |
| 1M8AB | 15596 | 1.7 | 0.343 | 61 | 59 | 58 | 0 | 0 | 0 | 0 |
| 1IOZA | 10051 | 2 | 0.339 | 162 | 156 | 0 | 0 | 0 | 0 | 153 |
| 2BJDA | 6398 | 1.27 | 0.337 | 90 | 87 | 81 | 0 | 0 | 0 | 0 |
| 1V6PB | 1381 | 0.87 | 0.333 | 62 | 55 | 51 | 0 | 0 | 0 | 0 |
| 1RB4B | 371 | 1.9 | 0.333 | 32 | 31 | 30 | 0 | 0 | 0 | 0 |
| 2PF5A | 6392, 6393 | 1.9 | 0.327 | 88 | 84 | 78 | 84 | 88 | 78 | 84 |
| 2SEMA | 5729 | 2.2 | 0.317 | 58 | 49 | 48 | 50 | 52 | 48 | 48 |
| 2IDSA | 16740 | 1 | 0.295 | 105 | 93 | 0 | 0 | 0 | 0 | 93 |
| 1EQTB | 16803 | 1.6 | 0.294 | 67 | 58 | 60 | 57 | 62 | 60 | 58 |
| 3DFR_ | 7200, | 1.7 | 0.286 | 162 | 147 | 0 | 0 | 0 | 0 | 137 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 7196, 7197, 7198, 7199 | | | | | | | | | |
| 1H0JB | 4966 | 1.9 | 0.283 | 60 | 54 | 58 | 0 | 59 | 0 | 0 |
| 1BAZC | 394, 395 | 1.9 | 0.283 | 46 | 41 | 40 | 0 | 0 | 0 | 0 |
| 1DCDB | 5249 | 2 | 0.278 | 36 | 35 | 26 | 0 | 32 | 0 | 35 |
| 1OKSA | 6568, 6569 | 1.8 | 0.268 | 50 | 48 | 0 | 0 | 0 | 0 | 47 |
| 1R1TA | 4128, 4306 | 1.7 | 0.262 | 98 | 96 | 0 | 0 | 97 | 0 | 93 |
| 2GITE | 5784, 5169, 3078 | 1.7 | 0.260 | 100 | 92 | 97 | 0 | 0 | 0 | 0 |
| 1W7ZC | 397 | 1.67 | 0.258 | 31 | 26 | 23 | 0 | 0 | 0 | 0 |
| 2IDUA | 16741 | 0.95 | 0.257 | 104 | 95 | 0 | 94 | 99 | 89 | 95 |
| 1UWXA | 1639 | 2.2 | 0.254 | 57 | 54 | 50 | 0 | 0 | 0 | 0 |
| 1HC9A | 8, 4195, 5006, 5024 | 1.8 | 0.243 | 74 | 63 | 68 | 0 | 0 | 0 | 0 |
| 1CM9B | 4914, 4852 | 2.1 | 0.243 | 66 | 60 | 63 | 0 | 0 | 0 | 58 |
| 1HC9B | 15130 | 1.8 | 0.243 | 74 | 0 | 70 | 0 | 74 | 0 | 0 |
| 1YU7X | 4428 | 1.5 | 0.239 | 64 | 57 | 58 | 0 | 0 | 0 | 56 |
| 1ZNIC | 1632, 1585, 554, 1344 | 1.5 | 0.238 | 21 | 20 | 20 | 0 | 0 | 0 | 0 |
| 2F91B | 5274 | 1.2 | 0.229 | 32 | 29 | 29 | 0 | 0 | 0 | 0 |
| 2ALGA | 16294 | 2.3 | 0.228 | 92 | 77 | 69 | 0 | 0 | 0 | 0 |
| 1PZ4A | 16662 | 1.35 | 0.224 | 113 | 106 | 104 | 0 | 109 | 100 | 102 |
| 2IE2C | 17162, 17163 | 1.7 | 0.223 | 212 | 194 | 0 | 0 | 191 | 161 | 174 |
| 1CLVI | 4490, 4404 | 2 | 0.219 | 32 | 27 | 29 | 0 | 0 | 0 | 0 |
| 2GJ2A | 7099 | 2.35 | 0.212 | 79 | 77 | 74 | 0 | 78 | 68 | 74 |
| 1PHT_ | 16448 | 2 | 0.212 | 83 | 0 | 0 | 46 | 69 | 46 | 61 |
| 2CA5B | 15214 | 2.1 | 0.212 | 61 | 50 | 0 | 0 | 16 | 10 | 39 |
| 1WKXA | 6123 | 1.7 | 0.209 | 43 | 31 | 30 | 0 | 0 | 0 | 0 |
| 1LP1B | 1120, 4324 | 2.3 | 0.207 | 54 | 12 | 13 | 0 | 0 | 0 | 12 |
| 2FFGA | 15529 | 2.31 | 0.207 | 77 | 68 | 0 | 0 | 61 | 57 | 61 |
| 1LU0B | 4246 | 1.03 | 0.207 | 29 | 27 | 26 | 0 | 0 | 0 | 0 |
| 1PPEI | 199, 314, 2227, 2527 | 2 | 0.207 | 29 | 3 | 5 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1HSLB | 16204, 16205 | 1.89 | 0.206 | 238 | 215 | 169 | 221 | 216 | 184 | 204 |
| 1PGX_ | 2575 | 1.66 | 0.205 | 70 | 58 | 54 | 0 | 0 | 0 | 0 |
| 1OVHA | 915 | 1.95 | 0.201 | 162 | 152 | 139 | 0 | 0 | 0 | 152 |
| 1DFNA | 16254 | 1.9 | 0.200 | 30 | 0 | 0 | 27 | 28 | 24 | 28 |
| 1CQMB | 16344 | 1.65 | 0.198 | 98 | 88 | 88 | 0 | 0 | 0 | 88 |
| 2IGD_ | 15283 | 1.1 | 0.197 | 61 | 0 | 0 | 54 | 54 | 50 | 53 |
| 1FD3D | 4642 | 1.35 | 0.195 | 41 | 35 | 35 | 0 | 0 | 0 | 0 |
| 451C_ | 1333, 10133 | 1.6 | 0.195 | 82 | 75 | 73 | 0 | 0 | 0 | 69 |
| 1O5UA | 16006 | 1.83 | 0.188 | 88 | 82 | 83 | 82 | 87 | 84 | 82 |
| 1RPJA | 16984, 16982 | 1.8 | 0.188 | 288 | 273 | 0 | 261 | 288 | 263 | 273 |
| 1QG7A | 16142 | 2 | 0.179 | 62 | 58 | 62 | 58 | 62 | 62 | 57 |
| 1EZGB | 5323 | 1.4 | 0.179 | 82 | 78 | 74 | 0 | 0 | 0 | 76 |
| 2NWGB | 16143 | 2.07 | 0.176 | 64 | 57 | 61 | 57 | 61 | 61 | 55 |
| 2GSVB | 15350 | 1.9 | 0.175 | 66 | 63 | 61 | 0 | 65 | 65 | 63 |
| 1R69_ | 2539 | 2 | 0.174 | 63 | 60 | 58 | 0 | 0 | 0 | 0 |
| 2PSPB | 2384 | 1.95 | 0.170 | 105 | 88 | 94 | 0 | 0 | 0 | 0 |
| 3WRP_ | 17010 | 1.8 | 0.167 | 101 | 85 | 0 | 0 | 57 | 53 | 65 |
| 1OS3D | 1633 | 1.95 | 0.167 | 28 | 26 | 25 | 0 | 0 | 0 | 0 |
| 1QE6C | 280 | 2.35 | 0.167 | 67 | 61 | 62 | 0 | 0 | 0 | 0 |
| 1WTQA | 5905, 5908 | 1.7 | 0.167 | 64 | 60 | 0 | 0 | 0 | 0 | 59 |
| 1TUKA | 4977 | 1.12 | 0.164 | 67 | 62 | 60 | 0 | 0 | 0 | 61 |
| 3ERAB | 7211 | 1.7 | 0.161 | 62 | 53 | 52 | 0 | 0 | 0 | 51 |
| 2DGCA | 1396, 1397, 1398, 1399 | 2.2 | 0.159 | 49 | 48 | 48 | 0 | 0 | 0 | 0 |
| 1O82D | 4112 | 1.46 | 0.157 | 70 | 69 | 60 | 0 | 0 | 0 | 0 |
| 1C8CA | 4570 | 1.45 | 0.156 | 64 | 56 | 51 | 0 | 0 | 0 | 0 |
| 2CG7A | 15756 | 1.2 | 0.156 | 90 | 83 | 0 | 0 | 88 | 72 | 83 |
| 1BWOB | 2065, 4932, 4383 | 2.1 | 0.156 | 90 | 84 | 81 | 0 | 0 | 0 | 0 |
| 2H9EC | 4396 | 2.2 | 0.155 | 52 | 50 | 48 | 0 | 2 | 4 | 50 |
| 1OMYA | 7330 | 2 | 0.154 | 64 | 59 | 60 | 0 | 0 | 0 | 0 |
| 1MIDA | 15143 | 1.71 | 0.154 | 91 | 83 | 73 | 0 | 0 | 0 | 0 |
| 1KBAB | 1675 | 2.3 | 0.152 | 66 | 60 | 58 | 0 | 0 | 0 | 0 |
| 1L1DB | 6051 | 1.85 | 0.151 | 145 | 113 | 100 | 119 | 125 | 95 | 99 |
| 1YU8X | 15245 | 1.45 | 0.149 | 64 | 39 | 40 | 0 | 0 | 0 | 0 |
| 1NAQF | 15094 | 1.7 | 0.143 | 105 | 99 | 41 | 104 | 105 | 98 | 104 |
| 1GZZB | 2498, 4204, 15654 | 2.3 | 0.143 | 60 | 54 | 54 | 57 | 59 | 44 | 50 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2A3GA | 1442 | 2.25 | 0.143 | 21 | 19 | 20 | 0 | 0 | 0 | 0 |
| 1KX9B | 5094 | 1.65 | 0.143 | 102 | 98 | 95 | 0 | 0 | 0 | 92 |
| 1BNZA | 6050 | 2 | 0.141 | 64 | 51 | 47 | 0 | 0 | 0 | 0 |
| 1ICFJ | 5583 | 2 | 0.138 | 65 | 58 | 59 | 0 | 0 | 0 | 0 |
| 1YNRB | 10135 | 2 | 0.138 | 79 | 75 | 0 | 0 | 0 | 0 | 72 |
| 1GN0A | 17136 | 1.8 | 0.130 | 108 | 103 | 99 | 0 | 108 | 100 | 103 |
| 1AE3_ | 2039 | 2 | 0.128 | 86 | 75 | 72 | 0 | 0 | 0 | 0 |
| 1HB8B | 2049 | 2 | 0.116 | 86 | 82 | 80 | 0 | 0 | 0 | 0 |
| 1KTZA | 4411 | 2.15 | 0.116 | 82 | 67 | 74 | 0 | 0 | 0 | 67 |
| 2GM5D | 4269 | 2.1 | 0.115 | 114 | 79 | 72 | 78 | 81 | 74 | 79 |
| 1CY5A | 4661 | 1.3 | 0.113 | 92 | 86 | 86 | 0 | 84 | 12 | 86 |
| 1VKBA | 16380 | 1.9 | 0.112 | 147 | 130 | 132 | 0 | 141 | 116 | 118 |
| 1QVEA | 4918 | 1.54 | 0.111 | 126 | 114 | 105 | 0 | 0 | 0 | 108 |
| 2GOOB | 15956 | 2.2 | 0.107 | 85 | 76 | 79 | 53 | 66 | 57 | 53 |
| 1J1VA | 5200 | 2.1 | 0.106 | 91 | 82 | 72 | 85 | 91 | 87 | 86 |
| 1THQA | 6234 | 1.9 | 0.100 | 147 | 102 | 0 | 0 | 99 | 75 | 91 |
| 2BEMC | 17160 | 1.55 | 0.100 | 170 | 152 | 159 | 168 | 169 | 159 | 152 |
| 1ENFA | 16146 | 1.69 | 0.099 | 212 | 204 | 0 | 0 | 200 | 165 | 198 |
| 1BGF_ | 5997 | 1.45 | 0.097 | 124 | 111 | 0 | 96 | 104 | 85 | 111 |
| 1Q2UA | 17507 | 1.6 | 0.095 | 189 | 171 | 0 | 178 | 179 | 148 | 157 |
| 2UV0H | 6271 | 1.8 | 0.091 | 163 | 150 | 0 | 161 | 161 | 147 | 150 |
| 1AM7A | 16664 | 2.3 | 0.089 | 150 | 143 | 126 | 141 | 146 | 0 | 143 |
| 1VYKA | 17357 | 1.49 | 0.087 | 129 | 82 | 71 | 91 | 99 | 96 | 93 |
| 1JPST | 16838 | 1.85 | 0.087 | 200 | 179 | 0 | 177 | 178 | 66 | 160 |
| 1XIOA | 17064 | 2 | 0.084 | 217 | 0 | 0 | 129 | 132 | 97 | 134 |
| 2IM8A | 16113, 7227 | 2 | 0.084 | 120 | 95 | 104 | 0 | 94 | 90 | 75 |
| 2BDYA | 16940 | 1.61 | 0.083 | 276 | 248 | 0 | 0 | 0 | 0 | 0 |
| 1ASS_ | 5930 | 2.3 | 0.082 | 152 | 145 | 140 | 133 | 147 | 0 | 140 |
| 2Q2TA | 16059 | 2.3 | 0.072 | 293 | 249 | 0 | 236 | 265 | 235 | 249 |
| 1K82A | 5219 | 2.1 | 0.071 | 260 | 172 | 0 | 106 | 148 | 95 | 118 |
| 1CKUB | 2999 | 1.2 | 0.071 | 85 | 75 | 73 | 0 | 0 | 0 | 0 |
| 1PHP_ | 16464, 16451, 16447 | 1.65 | 0.063 | 394 | 372 | 0 | 359 | 363 | 325 | 366 |
| 2ASDA | 16869 | 1.95 | 0.053 | 341 | 189 | 0 | 187 | 191 | 141 | 189 |
| 2GT8A | 17251 | 2 | 0.042 | 298 | 187 | 185 | 168 | 195 | 161 | 169 |
| 2ILNI | 5617 | 2 | 0.000 | 53 | 47 | 53 | 0 | 0 | 0 | 0 |

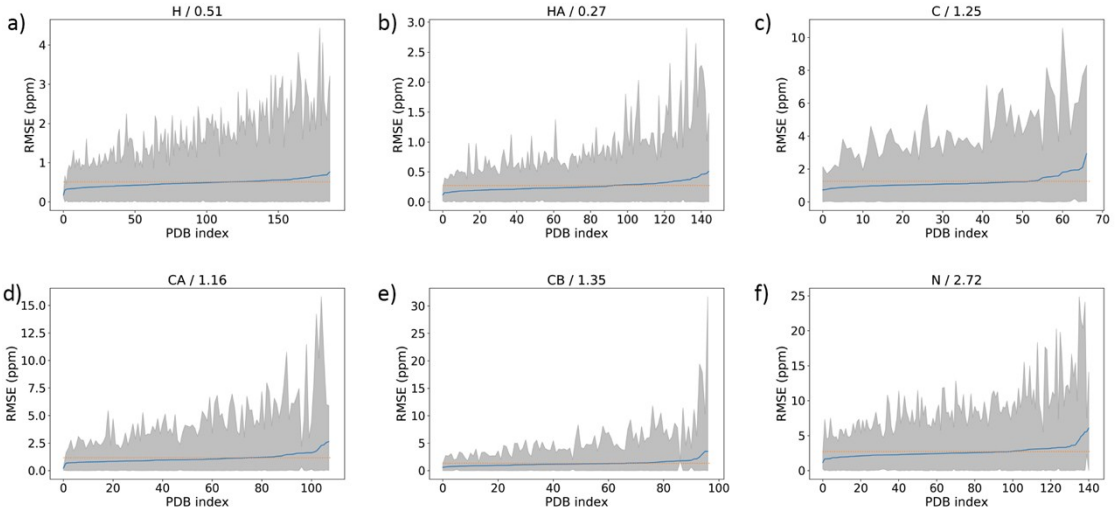
## Evaluation of SPARTA+, SHIFTX2 and UCBShift performance on Test datasets

**Table S4.** *Root mean square error (RMSE), mean squared error (MAE) and Pearson's correlation coefficients ($R^2$) for UCBShift in the curated and uncurated test dataset and its low homology subset.*

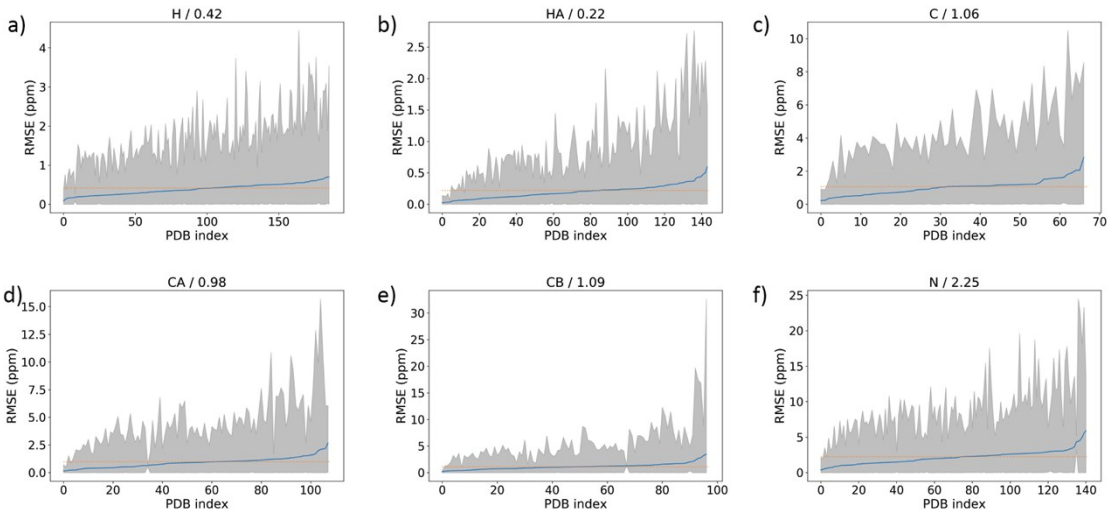
| Dataset | | Test | | | LH-Test | |
|---|---|---|---|---|---|---|
| **Atom Type** | **RMSE** | **MAE** | **$R^2$** | **RMSE** | **MAE** | **$R^2$** |
| **H** | 0.31±0.003 | 0.18±0.001 | 0.90 | 0.45±0.004 | 0.32±0.001 | 0.77 |
| **Hα** | 0.19±0.002 | 0.11±0.001 | 0.94 | 0.26±0.003 | 0.18±0.002 | 0.87 |
| **C'** | 0.84±0.01 | 0.48±0.004 | 0.93 | 1.14±0.01 | 0.81±0.007 | 0.86 |
| **Cα** | 0.81±0.01 | 0.43±0.004 | 0.99 | 1.09±0.01 | 0.73±0.005 | 0.97 |
| **Cβ** | 1.00±0.03 | 0.47±0.005 | 1.00 | 1.34±0.05 | 0.83±0.009 | 0.99 |
| **N** | 1.81±0.02 | 1.06±0.006 | 0.95 | 2.61±0.02 | 1.89±0.01 | 0.88 |
| **Dataset** | | **Test (Curated)** | | | **LH-Test (Curated)** | |
| **Atom Type** | **RMSE** | **MAE** | **$R^2$** | **RMSE** | **MAE** | **$R^2$** |
| **H** | 0.30±0.002 | 0.17±0.001 | 0.90 | 0.43±0.003 | 0.32±0.002 | 0.78 |
| **Hα** | 0.18±0.002 | 0.10±0.001 | 0.94 | 0.24±0.003 | 0.17±0.002 | 0.88 |
| **C'** | 0.77±0.008 | 0.44±0.004 | 0.94 | 1.13±0.01 | 0.81±0.006 | 0.86 |
| **Cα** | 0.76±0.009 | 0.41±0.003 | 0.99 | 1.04±0.01 | 0.71±0.006 | 0.98 |
| **Cβ** | 0.82±0.01 | 0.45±0.004 | 1.00 | 1.12±0.02 | 0.78±0.007 | 1.00 |
| **N** | 1.71±0.01 | 1.03±0.008 | 0.95 | 2.55±0.02 | 1.86±0.01 | 0.88 |

**Figure S2**. *Error analysis for SPARTA+, SHIFTX2 and UCBShift.* In each graph, the dotted line indicates the average RMSE over all the predicted residues. PDBs are sorted according to the prediction RMSE over that specific structure, and the grey region represents the minimum and maximum RMSE of a single residue in a protein for atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N atom types.

**SPARTA+**




**SHIFTX2**


a)
H / 0.42
b)
HA / 0.22
c)
C / 1.06
d)
CA / 0.98
e)
CB / 1.09
f)
N / 2.25
RMSE (ppm)
PDB index

**UCBShift**

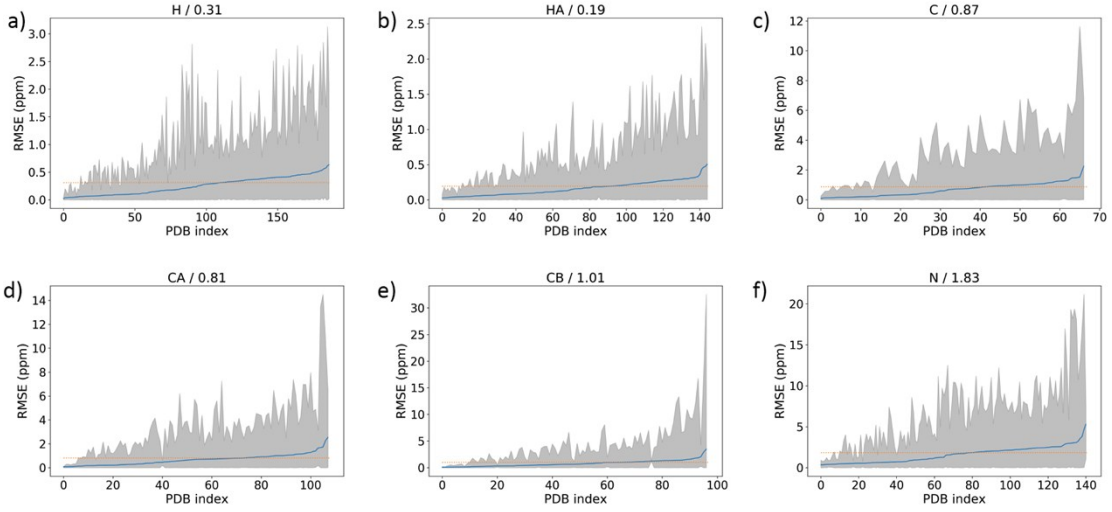
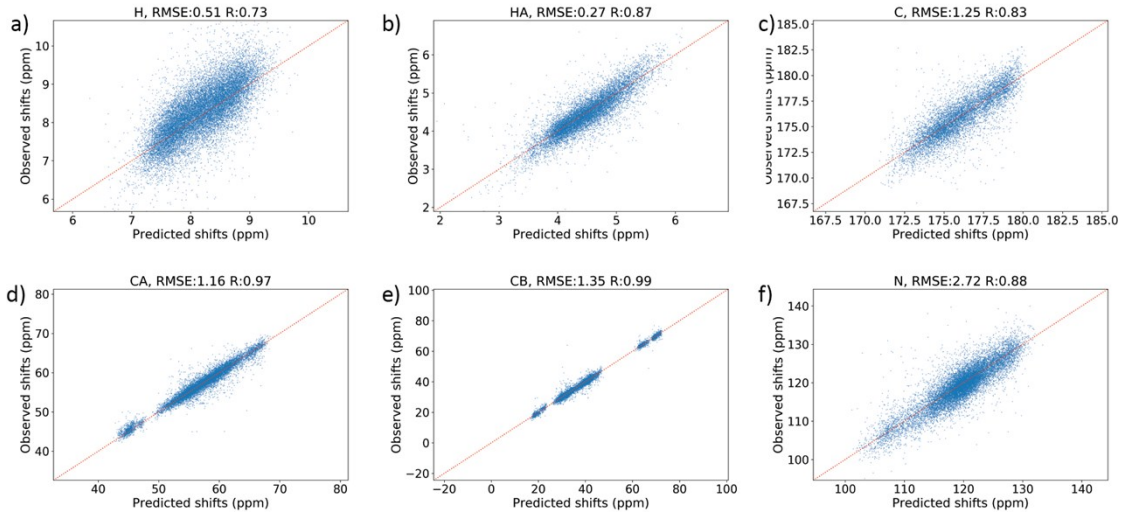
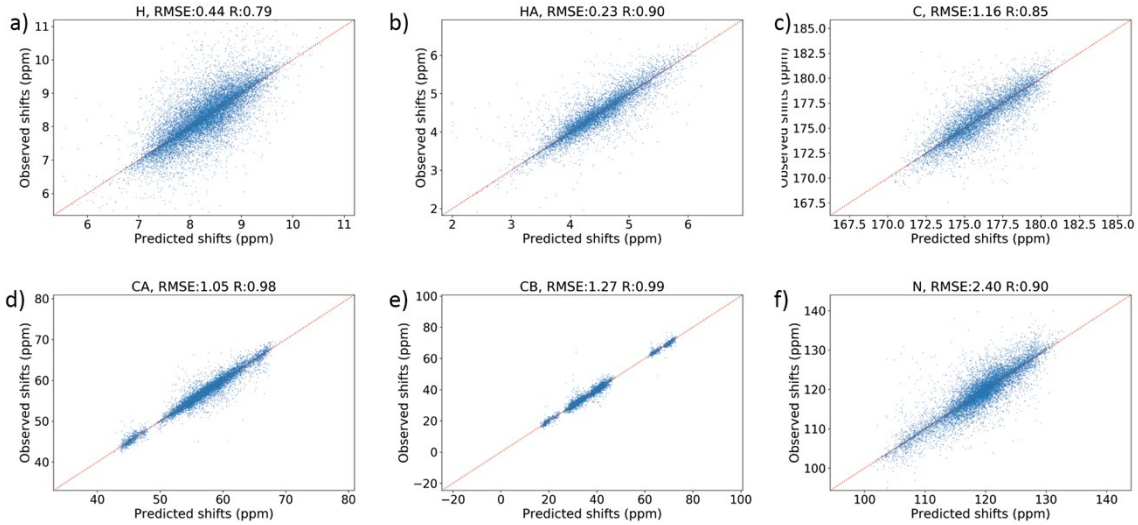

a)
H / 0.31
b)
HA / 0.19
c)
C / 0.87
d)
CA / 0.81
e)
CB / 1.01
f)
N / 1.83
RMSE (ppm)
PDB index

**Figure S3**. *Scatter plot for SPARTA+, SHIFTX2 and UCBShift predictions.* The predicted chemical shifts for six atom types plotted with the experimental shifts types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N atom types.

**SPARTA+**




**SHIFTX2**



**UCBShift**

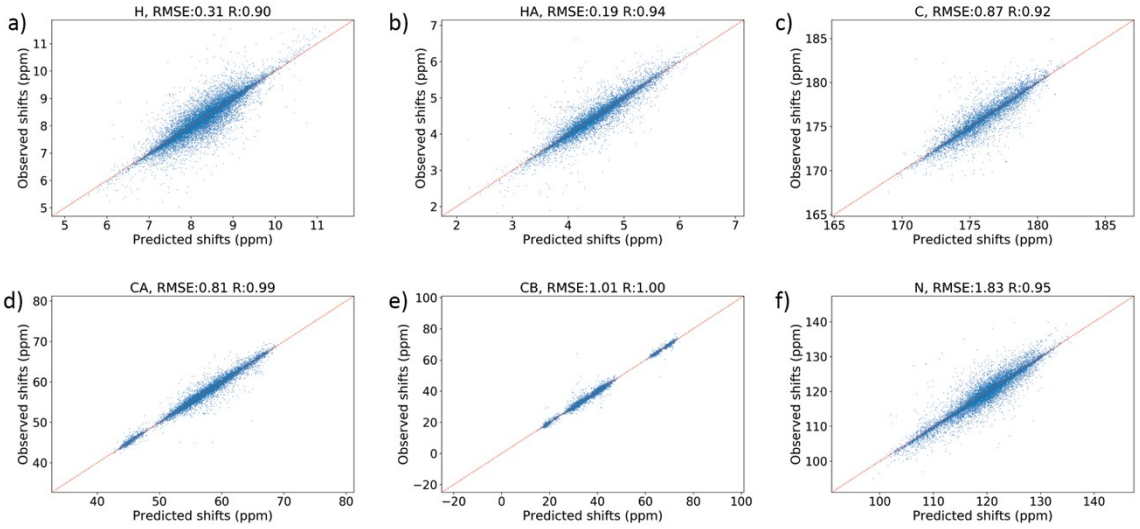
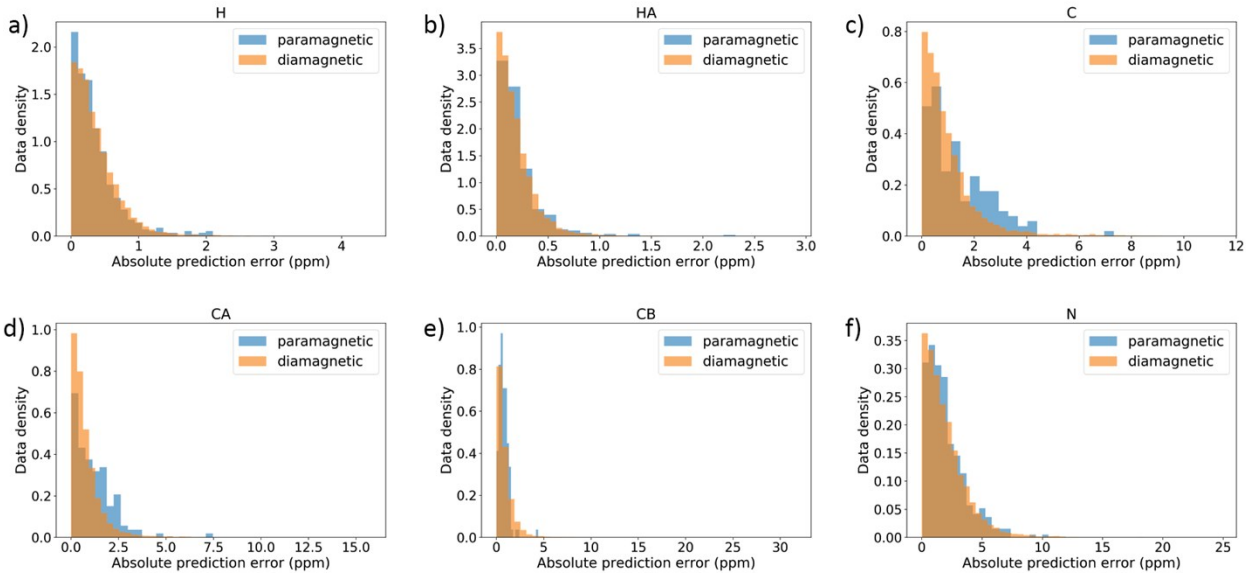



**Table S5.** Examples of SHIFTX2 that have information leakage from training to test set.

| Identifier in training | Identifier in testing | Sequence similarity | Structure RMSD (Å) |
|---|---|---|---|
| R014_3LZTA | A055_1YKYX | 100% | 0.77 |
| R114_1VDQA | A055_1YKYX | 100% | 0.41 |
| R020_1RUVA | A001_1KF3A | 100% | 0.20 |
| R006_2CPLA | A054_1CWCA | 98% * | 0.21 |
| R129_1ZJLA | A022_1ZJLA | 100% | 0 |
| R072_3RN3A | A001_1KF3A | 100% | 0.11 |

* Training (R006) is a sub-sequence of testing (A054) with 1 aa missing on N-terminus and 2 aa missing on C-terminus

**Figure S4**. *Comparison of absolute error distribution for paramagnetic proteins and diamagnetic proteins based on SPARTA+ predictions.* Atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N.



## Performance analysis of UCBShift-Y in comparison with SHIFTY+

**Figure S5.** *Difference between UCBShift-Y and SHIFTY+ for protein specific RMSEs for different atom types as a function of sequence identity.* Atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N.

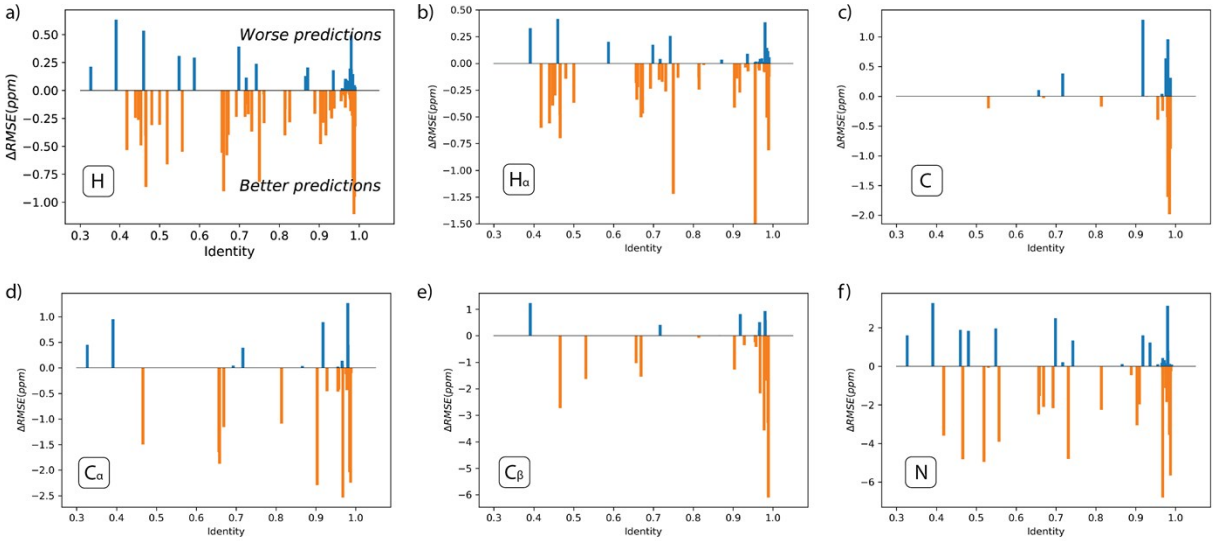
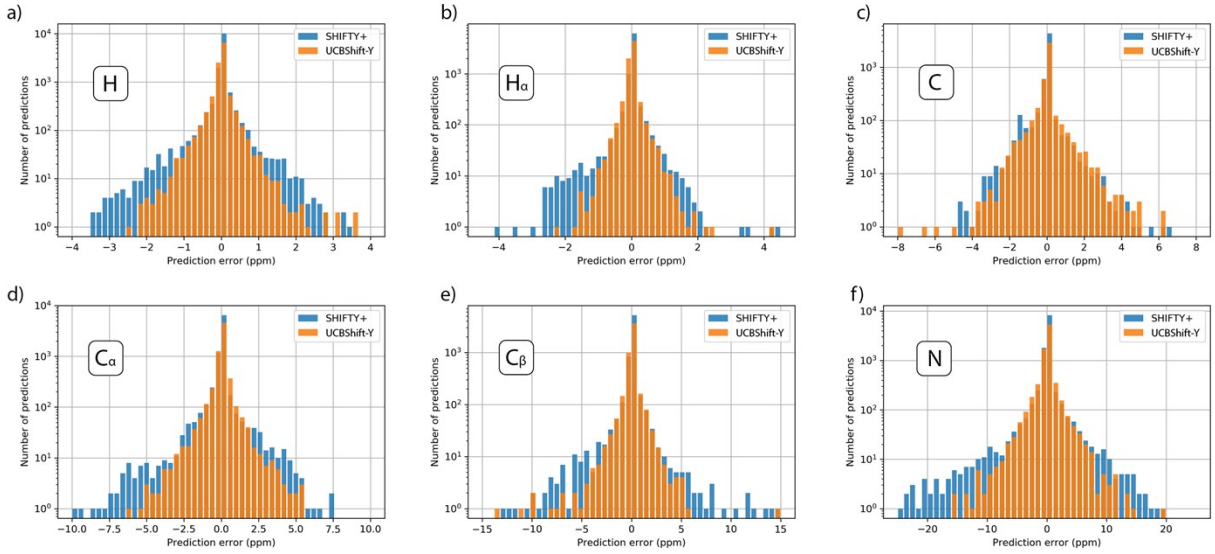


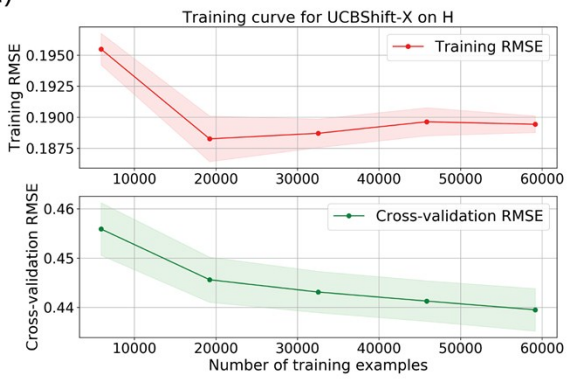

**Figure S6**. *Error distributions for UCBShift-Y and SHIFTY+ across all atom types.* This compares only the transfer prediction module of UCBShift (UCBShift-Y) which uses sequence and structural alignment.



## Learning curves for random forest models

**Figure S7**. *Learning curves for R1 and R2 in UCBShift-X module for hydrogen.*

a)




b)