Table complementing supplementary S4: Quantitative comparison of message lengths for stating amino acid (backbone + sidechain) dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models

This table provides a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) for stating dihedral angles (backbone + sidechain) of each of the twenty naturally occurring amino acids (aa). The the 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles. While we model the joint distributions of dihedral including the backbone, Dunbrack on the other hand only provides sidechain distributions conditional on the backbone. Hence ALA and GLY Dunbrack libraries are necessarily empty.

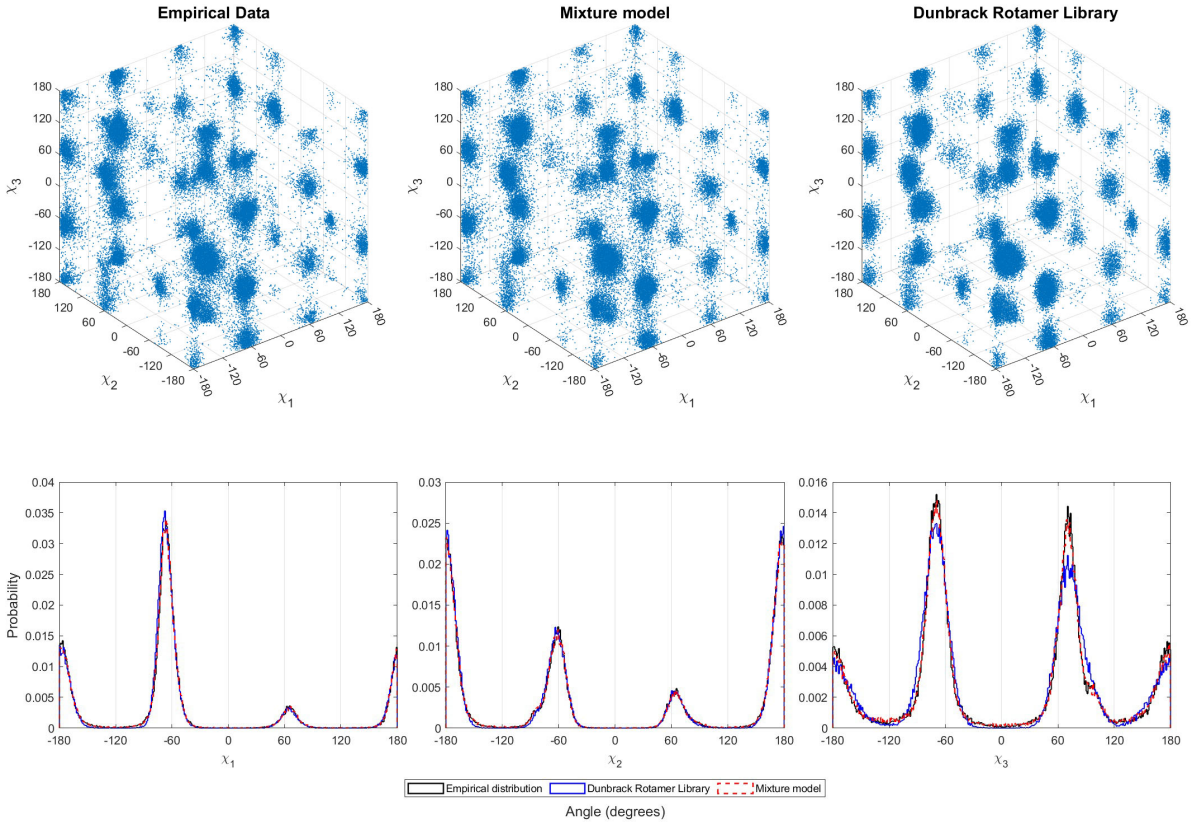
		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	$\text{Null}(X^{(\text{aa})})$	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(152;1,367)	6,939	4,926,340	4,933,279	14.4	(11,664;57,024)	950,779	6,587,654	7,538,433	21.9	8,483,696	24.7
ALA	334,111	(26;129)	779	2,509,435	2,510,214	7.5	(N/A;N/A)	N/A	N/A	N/A	N/A	4,122,880	12.3
GLY	294,278	(35;174)	900	2,829,700	2,830,600	9.6	(N/A;N/A)	N/A	N/A	N/A	N/A	3,631,346	12.3
VAL	274,596	(97;678)	3,706	2,918,937	2,922,643	10.6	(3,888;10,368)	192,834	4,163,843	4,356,677	15.9	5,082,710	18.5
GLU	238,682	(240;2,639)	11,801	5,095,773	5,107,574	21.4	(69,984;488,592)	8,509,192	6,189,754	14,698,946	61.6	7,363,250	30.8
ASP	227,558	(219;1,970)	9,252	3,801,033	3,810,284	16.7	(23,328;115,344)	2,062,259	4,632,995	6,695,254	29.4	5,616,063	24.7
SER	222,721	(110;769)	3,859	2,816,672	2,820,530	12.7	(3,888;10,368)	185,948	3,657,879	3,843,828	17.3	4,122,516	18.5
ILE	215,684	(130;1,169)	6,073	2,978,210	2,984,282	13.8	(11,664;57,024)	848,230	4,018,755	4,866,985	22.6	5,323,016	24.7
THR	212,562	(85;594)	3,076	2,488,418	2,491,493	11.7	(3,888;10,368)	187,511	3,295,033	3,482,543	16.4	3,934,475	18.5
LYS	195,868	(368;4,783)	19,991	4,962,216	4,982,208	25.4	(104,976;943,488)	12,526,749	5,878,978	18,405,727	94.0	7,250,945	37.0
ARG	188,400	(353;5,294)	24,041	5,201,624	5,225,666	27.7	(104,976;943,488)	13,518,806	5,758,992	19,277,798	102.3	8,136,897	43.2
PRO	177,534	(146;1,313)	8,925	1,974,713	1,983,637	11.2	(2,592;11,664)	225,394	3,195,740	3,421,135	19.3	4,381,486	24.7
ASN	162,196	(224;2,015)	9,194	2,865,575	2,874,770	17.7	(46,656;231,984)	4,025,549	3,472,766	7,498,315	46.2	4,002,949	24.7
PHE	153,192	(199;1,790)	8,552	2,516,999	2,525,552	16.5	(23,328;115,344)	1,940,884	3,077,402	5,018,286	32.8	3,780,733	24.7
GLN	136,703	(225;2,474)	10,776	2,947,251	2,958,026	21.6	(139,968;978,480)	16,100,149	3,615,527	19,715,676	144.2	4,217,236	30.8
TYR	134,950	(164;1,475)	6,884	2,237,744	2,244,627	16.6	(23,328;115,344)	1,970,652	2,718,877	4,689,528	34.8	3,330,526	24.7
HIS	89,382	(188;1,691)	7,605	1,593,555	1,601,160	17.9	(46,656;231,984)	3,818,112	1,928,561	5,746,674	64.3	2,205,921	24.7
MET	68,907	(209;2,298)	10,259	1,425,449	1,435,708	20.8	(34,992;243,648)	3,657,582	1,752,132	5,409,714	78.5	2,125,755	30.8
TRP	56,696	(154;1,385)	6,406	954,385	960,791	16.9	(46,656;231,984)	3,547,218	1,197,638	4,744,855	83.7	1,399,240	24.7
CYS	46,435	(72;503)	2,436	577,807	580,243	12.5	(3,888;10,368)	164,549	747,043	911,592	19.6	859,501	18.5

Table complementing supplementary S5: Quantitative comparison of message lengths for stating amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models

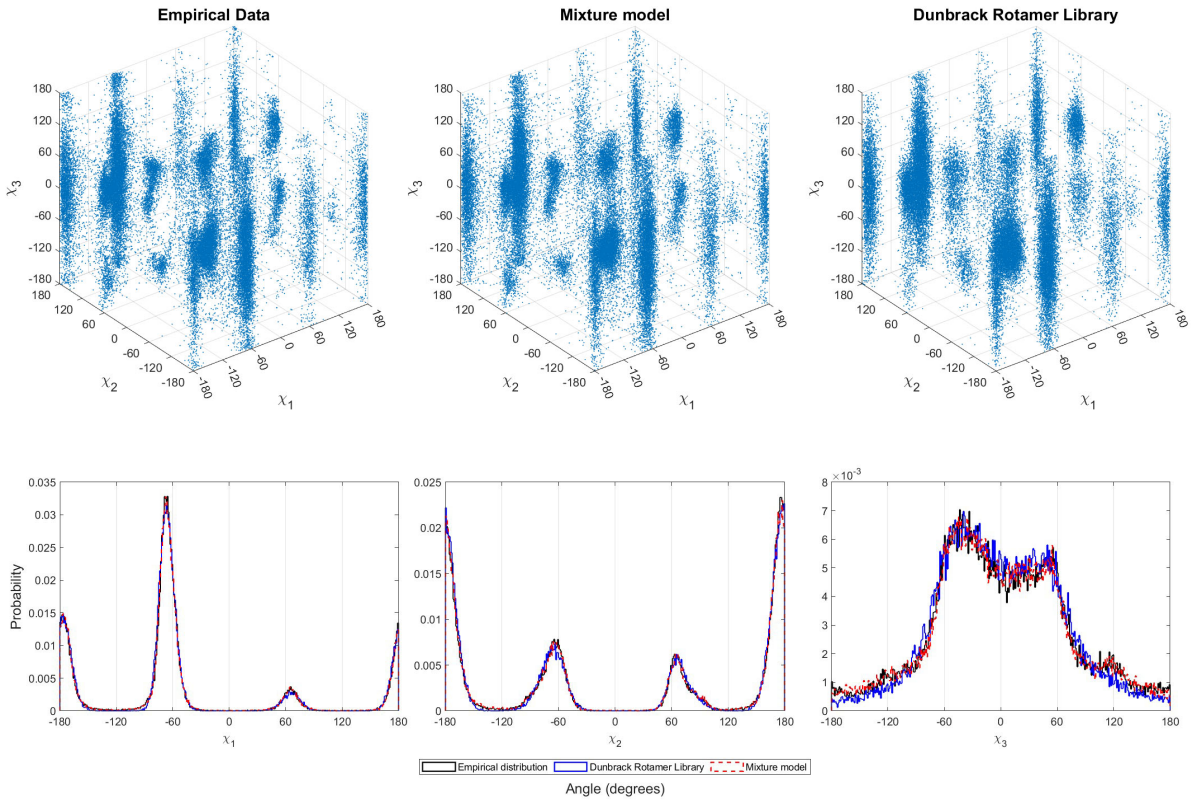
This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) to state only sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle\phi, \psi\rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	$\text{Null}(X^{(\text{aa})})$	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(152;1,367)	3,957	2,405,711	2,409,669	7.0	(11,664;57,024)	950,774	5,900,150	6,850,924	19.9	4,241,848	12.3
ALA	334,111	(26;129)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLY	294,278	(35;174)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	274,596	(97;678)	1,692	975,607	977,299	3.6	(3,888;10,368)	192,829	3,614,651	3,807,480	13.9	1,694,237	6.2
GLU	238,682	(240;2,639)	7,241	3,347,665	3,354,906	14.1	(69,984;488,592)	8,509,187	5,712,390	14,221,577	59.6	4,417,950	18.5
ASP	227,558	(219;1,970)	5,003	2,056,298	2,061,301	9.1	(23,328;115,344)	2,062,254	4,177,879	6,240,133	27.4	2,808,032	12.3
SER	222,721	(110;769)	1,696	994,203	995,899	4.5	(3,888;10,368)	185,943	3,212,437	3,398,380	15.3	1,374,172	6.2
ILE	215,684	(130;1,169)	3,481	1,519,622	1,523,103	7.1	(11,664;57,024)	848,225	3,587,387	4,435,612	20.6	2,661,508	12.3
THR	212,562	(85;594)	1,418	838,398	839,817	4.0	(3,888;10,368)	187,506	2,869,909	3,057,414	14.4	1,311,492	6.2
LYS	195,868	(368;4,783)	13,295	3,442,340	3,455,635	17.6	(104,976;943,488)	12,526,744	5,487,242	18,013,986	92.0	4,833,963	24.7
ARG	188,400	(353;5,294)	18,048	3,739,597	3,757,645	19.9	(104,976;943,488)	13,518,801	5,382,192	18,900,993	100.3	5,812,069	30.8
PRO	177,534	(146;1,313)	5,524	910,641	916,165	5.2	(2,592;11,664)	225,389	2,840,672	3,066,061	17.3	2,190,743	12.3
ASN	162,196	(224;2,015)	4,843	1,559,593	1,564,435	9.6	(46,656;231,984)	4,025,544	3,148,374	7,173,918	44.2	2,001,474	12.3
PHE	153,192	(199;1,790)	4,767	1,349,699	1,354,466	8.8	(23,328;115,344)	1,940,879	2,771,018	4,711,896	30.8	1,890,366	12.3
GLN	136,703	(225;2,474)	6,544	1,921,179	1,927,724	14.1	(139,968;978,480)	16,100,144	3,342,121	19,442,265	142.2	2,530,342	18.5
TYR	134,950	(164;1,475)	3,750	1,197,881	1,201,631	8.9	(23,328;115,344)	1,970,647	2,448,977	4,419,623	32.8	1,665,263	12.3
HIS	89,382	(188;1,691)	4,011	867,532	871,543	9.8	(46,656;231,984)	3,818,107	1,749,797	5,567,904	62.3	1,102,960	12.3
MET	68,907	(209;2,298)	6,489	905,273	911,762	13.2	(34,992;243,648)	3,657,577	1,614,318	5,271,895	76.5	1,275,453	18.5
TRP	56,696	(154;1,385)	3,507	529,675	533,183	9.4	(46,656;231,984)	3,547,212	1,084,246	4,631,458	81.7	699,620	12.3
CYS	46,435	(72;503)	1,065	200,506	201,571	4.3	(3,888;10,368)	164,544	654,173	818,717	17.6	286,500	6.2

Figure complementing supplementary S6: Qualitative comparison of model fit for methionine(MET) and glutamine(GLN) sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models



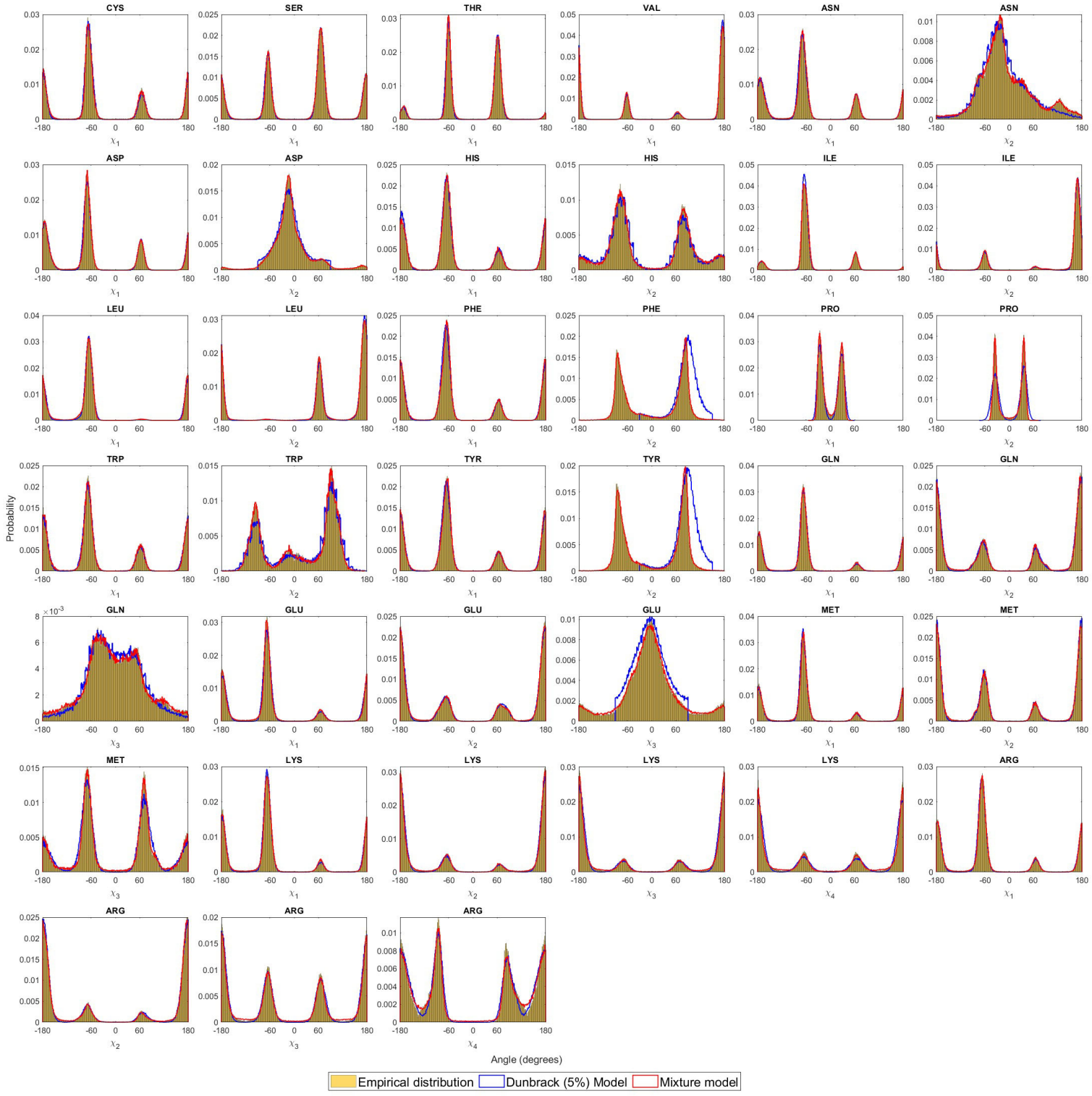
(a) Methionine (MET)



(b) Glutamine (GLN)

(a) The projection, into the sidechain (χ_1, χ_2, χ_3) space (unwrapped), of 50,000 randomly sampled points (vector of dihedral angles) for the amino acid Methionine (MET) from MML mixture model (first row, center), of the same number of points from the Dunbrack model (first row, right), and of the observed (empirical) distribution of the same angles (first row, left) from PDBHighRes. In the plots of the second row, the same data is visualized differently over three separate plots, with each of the three sidechain dihedral angles as x-axis (unwrapped), with y-axis showing the corresponding relative probabilities (in a 1° intervals). (b) The third and fourth rows plots are similar to first and second, respectively, but for the *non-rotameric* amino acid, Glutamine (GLN).

Figure complementing supplementary S7:Qualitative comparison of model fit across all amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models



Fidelity of the inferred MML mixure models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 1) from MML-derived mixture models and Dunbrack (5% smoothed) library, and compared to the empirical distribution.