

Supplementary material

Performance of Conformal Prediction in Capturing Aleatoric Uncertainty

Models' and conformal predictors' performance The models' accuracy and the conformal predictors' coverage, SSC, and mean set size (\bar{w}) on all four datasets are presented in Table A1 to Table A4.

Distribution of prediction set sizes Figure A1 shows the prediction set size distribution of the three conformal predictors.

Coverage at different prediction set sizes We show how coverage changes with increasing set sizes in Figure A2.

Impact of prevalence of larger prediction set sizes on correlation analysis We show how the correlation between prediction set sizes and class overlap changes with the prevalence of larger set sizes in Figure A3. We follow an incremental approach to see the effect of larger set sizes on the correlation analysis. We start with prediction sets with size ≤ 2 and incrementally add sets with larger sizes. The Spearman's rank correlation coefficient of prediction sets that contain larger set sizes shows improvement, albeit small.

Similarity between conformal prediction sets and human annotation We assess the similarity between the prediction sets of the conformal predictors and human annotations using the metrics precision, recall, subset-accuracy (S. Acc), and Hamming loss. The results on all the datasets are presented in Tables A5 to A8.

Correlation between conformal prediction sets and human annotation entropy We present the results of the correlation analysis between prediction sets and human annotation entropy on the CIFAR-10H and FER+ datasets in Table A9.

Expected Calibration Error Table A10 presents the Expected Calibration Error (ECE) using M=15 bins for all the models.

Table A1. Models' and conformal predictors' performance on the CIFAR-10H dataset.

Models	Accuracy	LAC			APS			RAPS		
		Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}
ResNet18	0.930	0.944	0.894	1.040	0.977	0.953	1.357	0.978	0.952	1.361
ResNet34	0.933	0.953	0.833	1.062	0.979	0.953	1.369	0.976	0.950	1.304
ResNet50	0.936	0.953	0.500	1.050	0.974	0.939	1.263	0.975	0.959	1.285
VGG-16	0.941	0.955	0.853	1.049	0.982	0.930	1.328	0.979	0.882	1.263
VGG-19	0.939	0.951	0.000	1.039	0.976	0.907	1.264	0.973	0.895	1.207
DenseNet121	0.940	0.942	0.000	1.006	0.979	0.938	1.395	0.975	0.947	1.274
DenseNet161	0.940	0.944	0.000	1.011	0.977	0.941	1.347	0.976	0.927	1.309
MobileNet-v2	0.939	0.951	0.000	1.030	0.984	0.953	1.425	0.982	0.966	1.370

Table A2. Models' and conformal predictors' performance on the MLRSNet dataset.

Models	Accuracy	LAC			APS			RAPS		
		Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}
ResNet18	0.927	0.951	0.000	1.081	0.984	0.889	1.575	0.985	0.900	1.624
ResNet34	0.923	0.951	0.831	1.090	0.982	0.500	1.500	0.982	0.750	1.526
ResNet50	0.918	0.949	0.780	1.113	0.978	0.667	1.504	0.978	0.667	1.508
VGG-16	0.872	0.950	0.750	1.613	0.955	0.833	1.798	0.963	0.887	2.059
VGG-19	0.844	0.955	0.889	2.106	0.958	0.875	2.352	0.974	0.953	3.386
DenseNet121	0.933	0.951	0.000	1.059	0.984	0.000	1.513	0.985	0.500	1.547
DenseNet161	0.946	0.956	0.000	1.026	0.985	0.928	1.320	0.984	0.905	1.316
MobileNet-v2	0.860	0.953	0.750	1.635	0.978	0.937	2.596	0.981	0.917	2.859

Table A3. Models' and conformal predictors' performance on the FER+ dataset.

Models	Accuracy	LAC			APS			RAPS		
		Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}
ResNet18	0.825	0.941	0.818	1.561	0.961	0.920	2.250	0.960	0.920	2.233
ResNet34	0.839	0.939	0.934	1.484	0.958	0.936	2.085	0.958	0.933	2.138
ResNet50	0.838	0.940	0.857	1.493	0.959	0.800	2.027	0.961	0.857	2.148
VGG-16	0.835	0.946	0.875	1.530	0.970	0.952	2.274	0.970	0.951	2.318
VGG-19	0.823	0.943	0.935	1.561	0.964	0.875	2.222	0.966	0.923	2.316
DenseNet121	0.832	0.945	0.833	1.521	0.965	0.833	2.199	0.967	0.857	2.273
DenseNet161	0.836	0.943	0.938	1.496	0.962	0.932	2.208	0.964	0.929	2.283
MobileNet-v2	0.827	0.944	0.940	1.563	0.965	0.952	2.237	0.966	0.952	2.311

Table A4. Models' and conformal predictors' performance on the ImageNet-Real dataset.

Models	Accuracy	LAC			APS			RAPS		
		Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}	Coverage	SSC	\bar{w}
ResNet18	0.725	0.902	0.500	2.994	0.939	0.000	12.680	0.939	0.000	12.712
ResNet34	0.759	0.899	0.681	2.157	0.940	0.000	8.894	0.940	0.500	8.943
ResNet50	0.785	0.903	0.000	1.760	0.946	0.000	7.768	0.946	0.000	7.892
VGG-16	0.745	0.902	0.000	2.461	0.943	0.000	10.690	0.943	0.000	10.587
VGG-19	0.751	0.903	0.000	2.355	0.946	0.000	10.764	0.946	0.000	10.692
DenseNet121	0.770	0.902	0.000	2.030	0.946	0.000	8.644	0.946	0.000	8.702
DenseNet161	0.796	0.902	0.000	1.633	0.942	0.000	6.053	0.942	0.000	6.034
MobileNet-v2	0.744	0.902	0.688	2.496	0.940	0.000	9.539	0.941	0.000	9.730

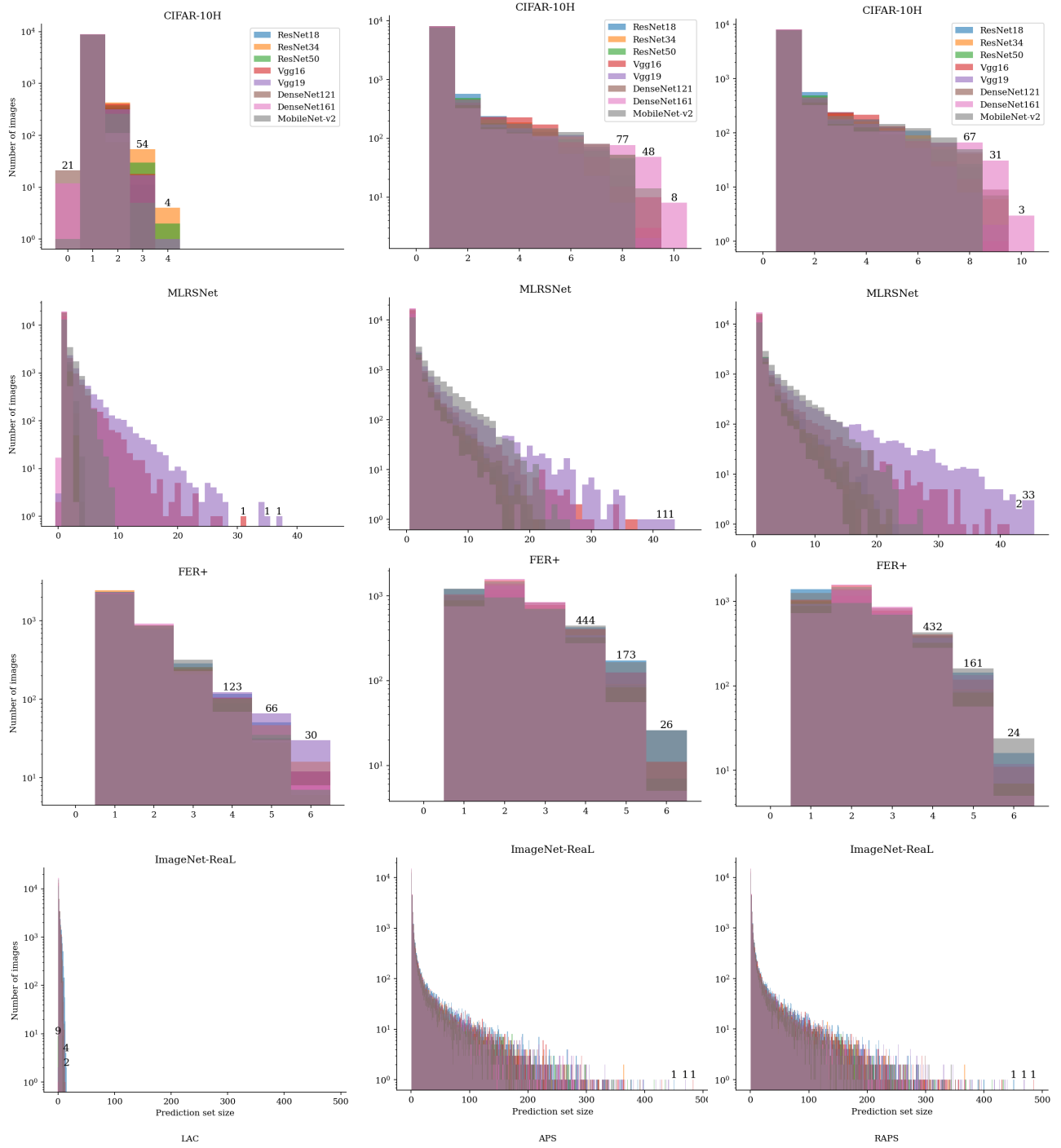


Figure A1. Distribution of prediction set sizes. The y-axis is log-scaled for easier visualization of the imbalanced counts. The count of the largest three prediction set sizes is shown on top of the bars. The three columns, from left to right, represent LAC, APS, and RAPS, respectively.

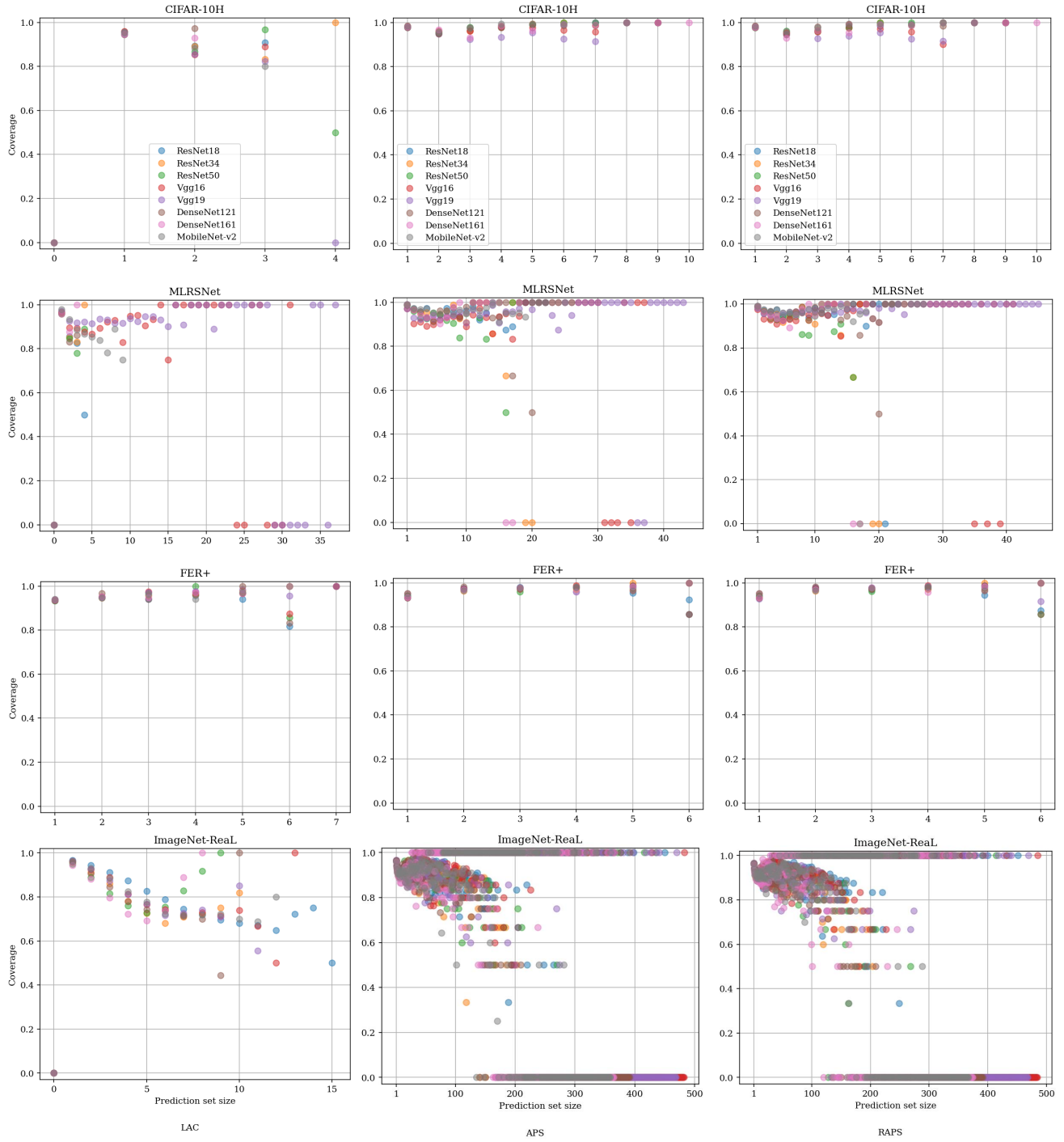


Figure A2. Coverage at different prediction set sizes. The X-axis shows prediction set sizes, and the Y-axis shows the coverage. The three columns, from left to right, represent LAC, APS, and RAPS, respectively.

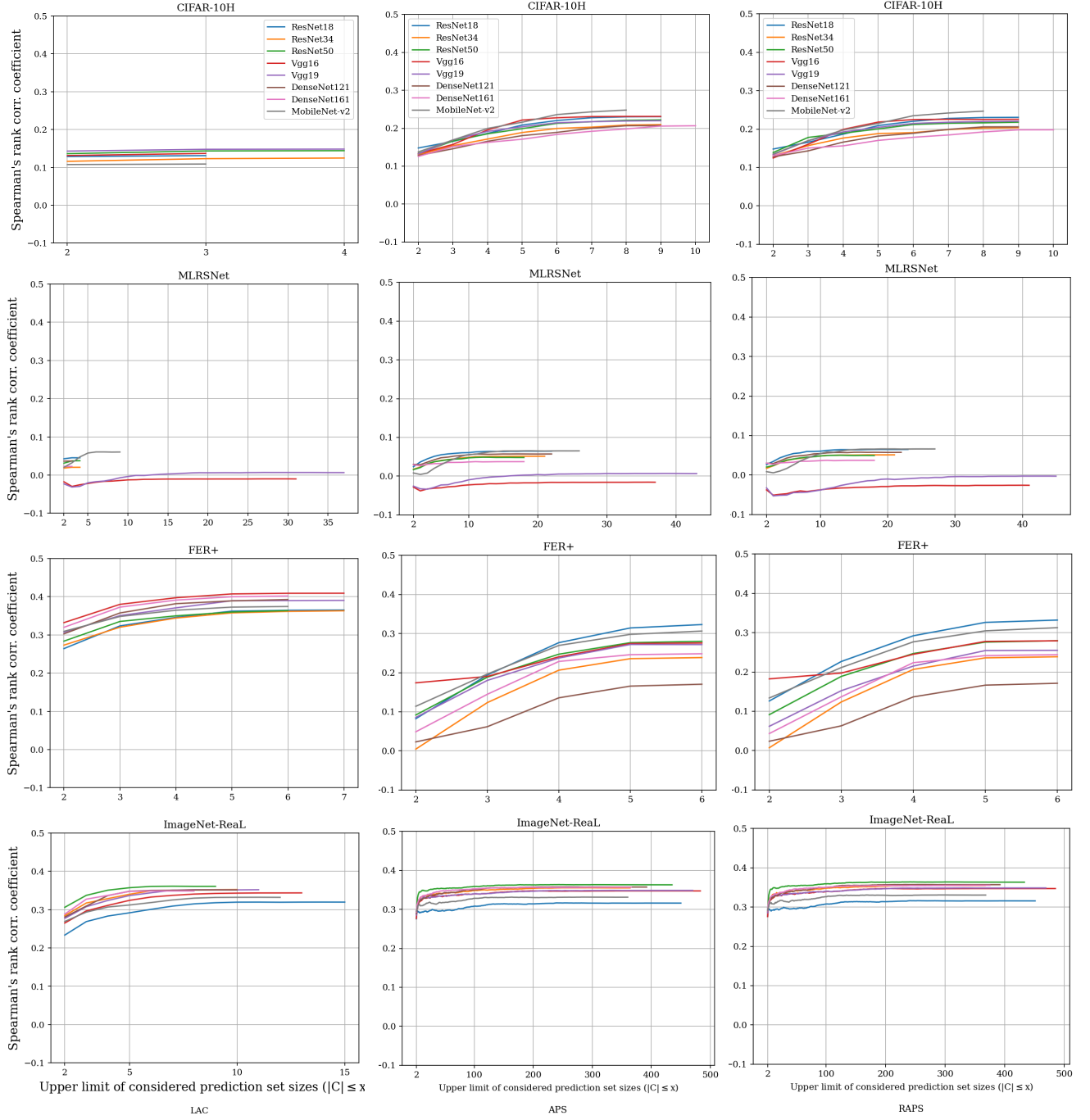


Figure A3. Spearman's rank correlation coefficient, r_s , $p < .001$, between prediction set sizes and class overlap with an increased prevalence of larger prediction sets. The X-axis shows the upper limit of the considered prediction set sizes ($|C| \leq x$), and the Y-axis shows Spearman's rank correlation coefficient. The three columns, from left to right, represent LAC, APS, and RAPS, respectively.

Table A5. Similarity between conformal predictors and human annotations on the CIFAR-10H dataset

LAC				APS				RAPS			
Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)
0.967	0.661	0.099	0.426	0.935	0.705	0.107	0.419	0.935	0.705	0.107	0.419
0.966	0.667	0.099	0.429	0.936	0.707	0.109	0.415	0.941	0.702	0.107	0.418
0.971	0.667	0.098	0.431	0.945	0.704	0.105	0.423	0.950	0.699	0.103	0.426
0.972	0.667	0.098	0.429	0.943	0.707	0.105	0.422	0.948	0.703	0.103	0.424
0.973	0.665	0.098	0.430	0.955	0.694	0.101	0.429	0.954	0.695	0.101	0.428
0.974	0.656	0.099	0.428	0.951	0.697	0.105	0.425	0.952	0.696	0.104	0.425
0.974	0.658	0.099	0.428	0.943	0.703	0.111	0.421	0.949	0.697	0.107	0.422
0.975	0.664	0.098	0.431	0.939	0.709	0.107	0.421	0.939	0.709	0.107	0.422

Table A6. Similarity between conformal predictors and human annotations on the MLRSNET dataset

LAC				APS				RAPS			
Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)
0.927	0.688	0.181	0.345	0.901	0.711	0.198	0.228	0.905	0.710	0.196	0.230
0.931	0.665	0.173	0.351	0.889	0.720	0.192	0.219	0.884	0.722	0.197	0.218
0.936	0.619	0.172	0.366	0.921	0.693	0.190	0.245	0.918	0.697	0.188	0.242
0.941	0.655	0.166	0.352	0.900	0.728	0.180	0.209	0.907	0.709	0.184	0.206
0.930	0.677	0.174	0.370	0.922	0.709	0.195	0.223	0.919	0.715	0.192	0.221
0.931	0.682	0.160	0.351	0.899	0.698	0.201	0.208	0.892	0.697	0.199	0.207
0.932	0.690	0.151	0.361	0.891	0.755	0.175	0.211	0.899	0.744	0.187	0.213
0.933	0.674	0.163	0.373	0.881	0.699	0.202	0.201	0.895	0.723	0.201	0.198

Table A7. Similarity between conformal predictors and human annotations on the FER+ dataset

LAC				APS				RAPS			
Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)
0.918	0.710	0.146	0.364	0.753	0.775	0.196	0.234	0.785	0.766	0.185	0.267
0.934	0.698	0.144	0.365	0.756	0.762	0.186	0.214	0.754	0.763	0.187	0.213
0.933	0.707	0.141	0.376	0.758	0.776	0.182	0.239	0.760	0.776	0.181	0.241
0.934	0.717	0.137	0.384	0.756	0.795	0.186	0.249	0.761	0.794	0.185	0.253
0.928	0.713	0.144	0.374	0.760	0.782	0.186	0.238	0.740	0.788	0.192	0.224
0.928	0.712	0.139	0.377	0.740	0.784	0.190	0.217	0.739	0.784	0.191	0.217
0.936	0.713	0.136	0.386	0.727	0.785	0.190	0.209	0.723	0.786	0.192	0.205
0.918	0.716	0.143	0.375	0.758	0.782	0.194	0.242	0.765	0.781	0.191	0.252

Table A8. Similarity between conformal predictors and human annotations on the ImageNet-Real dataset

LAC				APS				RAPS			
Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)	Precision	Recall	S. Acc	Hamming loss (\downarrow)
0.571	0.892	0.002	0.370	0.568	0.934	0.012	0.417	0.567	0.934	0.012	0.416
0.670	0.885	0.001	0.481	0.631	0.930	0.008	0.480	0.629	0.931	0.008	0.478
0.736	0.882	0.001	0.559	0.678	0.931	0.007	0.529	0.679	0.931	0.007	0.529
0.628	0.887	0.002	0.434	0.604	0.933	0.010	0.453	0.603	0.934	0.010	0.452
0.643	0.886	0.002	0.452	0.611	0.934	0.010	0.462	0.612	0.934	0.010	0.462
0.684	0.884	0.001	0.494	0.632	0.934	0.008	0.479	0.632	0.934	0.008	0.479
0.763	0.876	0.001	0.593	0.705	0.921	0.005	0.552	0.702	0.923	0.005	0.548
0.625	0.887	0.002	0.430	0.604	0.930	0.008	0.452	0.599	0.932	0.009	0.447

Table A9. Spearman's rank correlation coefficient, r_s , $p < .001$, between conformal prediction set sizes and human annotation entropy.

Models	CIFAR-10H			FER+		
	LAC	APS	RAPS	LAC	APS	RAPS
ResNet18	0.140	0.255	0.255	0.404	0.354	0.365
ResNet34	0.142	0.233	0.228	0.413	0.272	0.272
ResNet50	0.160	0.246	0.240	0.413	0.309	0.310
VGG-16	0.153	0.254	0.246	0.457	0.297	0.301
VGG-19	0.162	0.240	0.241	0.429	0.296	0.276
DenseNet121	0.025	0.230	0.228	0.437	0.191	0.191
DenseNet161	0.058	0.228	0.220	0.454	0.277	0.272
MobileNet-v2	0.124	0.269	0.268	0.427	0.337	0.343

Table A10. Expected Calibration Error (ECE) with M = 15 bins.

Models	CIFAR-10H	MLRSNet	FER+	ImageNet-Real
ResNet18	0.021	0.013	0.032	0.021
ResNet34	0.026	0.023	0.035	0.027
ResNet50	0.023	0.030	0.026	0.032
VGG-16	0.016	0.100	0.037	0.020
VGG-19	0.022	0.119	0.053	0.020
DenseNet121	0.020	0.018	0.034	0.020
DenseNet161	0.021	0.019	0.038	0.047
MobileNet-v2	0.025	0.012	0.027	0.020