

Appendix: The Surprising Utility of Group Partitioning in Improving Conformal Prediction of Visual Classifiers under Distributional Shifts

Kowshik Thopalli, Vivek Narayanaswamy
Lawrence Livermore National Laboratory

Jayaraman J. Thiagarajan
Apple

A. Expanded Results of Figure 1

Table 1. CP results for SWINv2-B, trained on ImageNet and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$, are reported. We present results averaged across all corruption types and severity levels for various scoring functions and grouping methods, with the number of groups set to 6 for NLL grouping. The best and second best values are colored in green and red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.12	5.88	92.83	4.44	6.79	0.40	6.56	2.43
	Mondrian	90.79	14.55	93.04	12.1	5.92	0.14	5.23	3.14
	Class Clustered	89.91	7.63	92.55	5.73	6.36	0.33	5.28	2.14
	NLL	90.89	13.45	91.43	8.6	1.57	0.49	6.25	2.29
APS	No Grouping	90.19	31.67	90.61	24.03	2.23	0.39	5.07	2.43
	Mondrian	90.99	55.64	90.9	47.72	1.95	0.03	5.0	3.71
	Class Clustered	89.68	35.09	89.82	27.9	1.89	0.24	4.59	2.14
	NLL	90.31	35.84	90.33	23.86	0.49	0.41	4.89	1.71
RAPS	No Grouping	90.20	9.03	93.11	8.81	7.33	0.12	6.6	2.43
	Mondrian	90.91	20.78	92.43	20.66	4.73	0.0	5.15	2.93
	Class Clustered	90.31	17.79	92.44	17.68	5.66	0.0	5.17	2.79
	NLL	90.67	14.38	91.00	9.17	1.13	0.41	5.76	1.86
SAPS	No Grouping	90.16	8.84	93.1	8.58	7.37	0.16	6.56	2.57
	Mondrian	90.91	20.57	92.34	20.38	4.57	0.01	5.11	3.0
	Class Clustered	90.06	18.28	92.17	18.09	5.49	0.0	5.18	2.57
	NLL	90.53	14.35	90.80	9.18	0.93	0.39	5.3	1.86

B. Expanded Results on PACS

We present the results for all training and calibration/testing configurations outlined in Section 4.3.2 on the PACS dataset. Additionally, we report performance using both GMM and flow-based likelihood estimation models. Across all configurations, we find that likelihood-aware CP consistently achieves lower GroupCovGap while maintaining comparable interval sizes and yielding higher Singleton Hit Ratio scores.

Table 2. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$, and calibrated/tested on equal samples from $\{P \cup A\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood through a flow-based model.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.85	3.26	90.04	3.85	3.9	0.26	9.7	2.43
	Mondrian	91.10	2.13	90.12	2.43	3.3	0.35	3.11	1.57
	Class Clustered	93.45	3.71	92.88	3.96	3.09	0.0	5.97	3.29
	NLL	92.28	3.63	92.45	4.28	2.67	0.31	6.81	2.71
APS	No Grouping	90.60	3.25	89.78	3.84	4.0	0.26	9.57	2.43
	Mondrian	92.28	2.44	92.00	2.84	3.04	0.33	3.58	1.57
	Class Clustered	93.87	4.12	93.89	4.51	3.89	0.0	6.42	3.71
	NLL	91.35	3.59	92.04	4.24	2.48	0.29	6.13	2.29
RAPS	No Grouping	88.66	3.71	84.85	3.81	9.29	0.0	9.57	3.5
	Mondrian	90.01	2.42	87.85	2.57	5.95	0.14	3.26	1.57
	Class Clustered	93.79	3.65	92.79	3.77	2.79	0.0	5.41	2.79
	NLL	91.86	3.46	91.94	4.06	2.16	0.24	7.35	2.14
SAPS	No Grouping	88.58	3.69	85.43	3.85	8.55	0.0	8.89	3.71
	Mondrian	91.77	2.73	90.21	2.94	3.40	0.12	3.75	1.71
	Class Clustered	91.35	3.36	88.88	3.49	5.82	0.01	5.90	2.43
	NLL	91.18	3.47	90.57	4.05	2.96	0.24	7.73	2.14

Table 3. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$, and calibrated/tested on equal samples from $\{P \cup C\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood through a flow-based model.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.66	4.44	93.13	5.14	7.65	0.13	5.11	2.71
	Mondrian	91.49	3.59	93.42	4.13	6.76	0.16	1.50	1.86
	Class Clustered	92.38	4.56	93.99	5.12	6.04	0.05	3.91	3.29
	NLL	91.26	4.24	91.15	4.62	1.15	0.11	5.99	2.14
APS	No Grouping	90.66	4.4	93.09	5.07	7.46	0.13	5.0	3.07
	Mondrian	90.74	4.25	92.73	4.86	6.73	0.15	2.28	1.79
	Class Clustered	92.08	4.29	93.71	4.87	6.14	0.13	4.32	2.93
	NLL	90.74	4.28	90.18	4.61	1.54	0.1	6.41	2.21
RAPS	No Grouping	90.81	4.19	89.79	4.29	2.27	0.00	8.74	2.93
	Mondrian	91.64	3.61	91.29	3.79	1.96	0.03	2.09	1.93
	Class Clustered	91.11	3.65	91.28	3.87	2.66	0.03	2.38	2.36
	NLL	91.34	4.07	90.86	4.27	2.25	0.00	8.19	2.79
SAPS	No Grouping	89.47	4.24	88.28	4.36	3.42	0.00	8.97	2.86
	Mondrian	90.44	3.87	89.86	4.05	3.23	0.01	2.05	1.14
	Class Clustered	92.68	4.99	92.56	5.1	3.24	0.00	4.07	3.43
	NLL	91.56	4.3	90.97	4.51	1.89	0.00	8.42	2.57

Table 4. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$, and calibrated/tested on equal samples from $\{P \cup S\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood through a flow-based model.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.54	3.54	90.69	3.74	3.45	0.18	11.53	2.57
	Mondrian	90.43	2.93	91.0	3.06	3.21	0.19	1.62	1.64
	Class Clustered	90.43	3.91	91.81	4.05	4.55	0.01	9.00	3.5
	NLL	90.19	3.59	90.81	3.75	1.76	0.17	10.93	2.29
APS	No Grouping	90.05	3.57	90.88	3.75	3.11	0.17	10.77	2.21
	Mondrian	89.86	3.48	90.25	3.64	2.13	0.17	1.47	1.36
	Class Clustered	88.93	4.39	89.91	4.6	3.45	0.01	9.35	3.29
	NLL	90.99	3.68	91.59	3.84	2.26	0.16	10.81	3.14
RAPS	No Grouping	90.85	4.04	91.97	4.03	3.33	0.00	14.01	2.50
	Mondrian	90.43	3.76	90.73	3.72	1.20	0.00	1.56	1.36
	Class Clustered	89.82	4.57	90.88	4.66	3.68	0.00	14.21	3.07
	NLL	91.51	4.1	92.49	4.09	3.36	0.00	13.89	3.07
SAPS	No Grouping	90.15	4.07	90.94	4.16	2.7	0.00	13.09	2.21
	Mondrian	90.15	4.11	90.81	4.25	2.22	0.01	1.17	1.50
	Class Clustered	89.72	4.79	91.29	4.95	4.65	0.00	11.62	3.43
	NLL	90.52	4.15	91.17	4.18	2.44	0.00	12.63	2.86

Table 5. Conformal Prediction (CP) results for Resnet50, trained on $\{P \cup A\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup A \cup S\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood through a flow-based model.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.32	2.61	89.75	3.08	3.94	0.29	8.53	2.43
	Mondrian	90.49	2.23	90.53	2.65	3.09	0.36	1.68	1.43
	Class Clustered	90.92	3.19	91.27	3.61	3.73	0.02	2.86	3.57
	NLL	90.66	2.69	90.64	3.18	1.84	0.32	7.8	2.57
APS	No Grouping	90.11	2.64	89.88	3.12	3.39	0.29	7.5	1.86
	Mondrian	90.75	2.57	91.71	3.11	3.96	0.31	1.71	2.00
	Class Clustered	89.85	3.52	91.19	4.05	4.98	0.02	3.64	3.29
	NLL	91.05	2.74	90.82	3.21	1.59	0.28	8.08	2.86
RAPS	No Grouping	89.94	3.12	87.06	3.19	7.47	0.0	11.28	3.14
	Mondrian	90.45	2.62	89.82	2.87	2.35	0.12	1.29	1.86
	Class Clustered	90.58	3.79	89.97	4.01	3.1	0.01	4.73	3.0
	NLL	90.15	2.84	90.14	3.34	0.37	0.22	8.07	2.00
SAPS	No Grouping	90.15	3.06	88.41	3.30	5.04	0.0	9.62	2.86
	Mondrian	90.24	3.17	90.40	3.46	2.7	0.09	1.81	2.43
	Class Clustered	89.68	3.15	89.81	3.45	2.24	0.08	2.67	2.29
	NLL	89.64	3.00	89.37	3.54	1.22	0.22	8.39	2.43

Table 6. Conformal Prediction (CP) results for Resnet50, trained on $\{P \cup A \cup C\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup A \cup C \cup S\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood through a flow-based model.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.69	1.68	88.14	1.92	4.19	0.53	7.72	2.71
	Mondrian	89.84	1.64	88.46	1.85	4.07	0.50	2.51	1.71
	Class Clustered	90.43	2.59	89.00	2.81	3.91	0.05	3.12	3.0
	NLL	89.11	1.83	89.20	2.15	1.16	0.49	6.17	2.57
APS	No Grouping	89.61	1.74	88.66	2.02	3.21	0.52	7.77	2.57
	Mondrian	89.73	1.83	90.05	2.13	2.37	0.45	2.73	2.43
	Class Clustered	90.16	1.80	89.92	2.09	2.33	0.46	3.49	2.00
	NLL	89.18	1.88	89.0	2.21	1.11	0.48	6.54	3.0
RAPS	No Grouping	90.74	1.81	87.84	1.92	7.02	0.22	7.79	2.86
	Mondrian	90.78	1.88	88.98	1.95	4.68	0.24	2.01	2.43
	Class Clustered	89.77	2.72	87.98	2.82	4.61	0.02	5.09	3.0
	NLL	89.96	1.86	89.86	2.19	0.72	0.39	6.79	1.71
SAPS	No Grouping	89.61	1.90	87.29	2.06	6.12	0.30	8.0	2.43
	Mondrian	90.74	2.01	89.48	2.15	3.69	0.25	2.39	2.29
	Class Clustered	89.49	2.92	88.74	3.07	2.44	0.01	3.21	3.0
	NLL	89.18	1.97	89.32	2.32	1.14	0.39	5.85	2.29

Table 7. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup A\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood using a GMM.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.85	3.26	91.74	2.83	4.08	0.26	9.7	2.07
	Mondrian	91.10	2.13	92.99	1.91	4.54	0.35	3.11	1.86
	Class Clustered	91.86	3.27	93.16	3.04	4.1	0.02	5.01	3.43
	NLL	91.18	3.38	91.74	3.02	2.04	0.25	8.4	2.64
APS	No Grouping	90.43	3.22	91.60	2.80	4.08	0.27	9.42	2.29
	Mondrian	92.53	2.43	93.97	2.17	3.99	0.33	3.77	1.86
	Class Clustered	92.44	3.11	94.1	2.85	4.13	0.03	4.99	3.14
	NLL	91.60	3.4	92.62	3.05	2.85	0.25	8.08	2.71
RAPS	No Grouping	88.66	3.7	90.25	3.59	4.84	0.0	9.57	2.93
	Mondrian	90.09	2.48	92.38	2.40	4.92	0.12	2.85	2.29
	Class Clustered	90.01	2.44	92.45	2.35	5.14	0.13	2.51	1.86
	NLL	88.66	3.77	90.40	3.64	4.58	0.08	8.71	2.93
SAPS	No Grouping	88.83	3.73	90.41	3.62	4.96	0.0	8.79	2.93
	Mondrian	91.02	2.54	92.36	2.39	3.33	0.14	3.36	1.43
	Class Clustered	93.79	4.09	95.3	4.02	5.46	0.0	5.23	3.64
	NLL	89.25	3.65	90.87	3.50	4.39	0.09	8.63	2.00

Table 8. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup C\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood using a GMM.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.66	4.44	91.33	4.00	2.9	0.13	5.11	2.36
	Mondrian	91.49	3.59	91.81	3.27	2.76	0.16	1.50	1.57
	Class Clustered	91.56	4.48	92.26	4.14	3.22	0.04	4.55	3.29
	NLL	91.64	4.61	91.08	4.18	2.50	0.13	5.06	2.79
APS	No Grouping	90.74	4.41	91.54	3.99	3.05	0.14	4.8	1.93
	Mondrian	91.78	4.27	92.6	3.86	3.75	0.15	2.83	1.86
	Class Clustered	93.13	4.6	93.65	4.26	3.77	0.01	3.82	3.71
	NLL	91.56	4.58	91.73	4.16	2.47	0.14	4.98	2.5
RAPS	No Grouping	90.96	4.19	91.83	4.16	2.62	0.0	8.49	2.79
	Mondrian	91.49	3.59	92.07	3.49	2.65	0.03	2.46	2.00
	Class Clustered	91.49	3.92	92.31	3.86	2.65	0.0	4.51	2.93
	NLL	91.26	3.97	91.15	3.64	2.91	0.08	7.33	2.29
SAPS	No Grouping	89.92	4.27	90.80	4.2	3.58	0.0	8.89	3.36
	Mondrian	91.41	3.93	91.82	3.84	2.66	0.01	1.52	2.29
	Class Clustered	90.96	4.19	91.57	4.14	3.19	0.0	4.45	2.93
	NLL	90.07	4.11	89.63	3.78	2.52	0.07	7.73	1.43

Table 9. Conformal Prediction (CP) results for Resnet50, trained on $\{P\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup S\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood using a GMM.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.54	3.54	90.46	3.46	2.16	0.18	11.53	2.93
	Mondrian	90.43	2.93	91.97	2.91	4.13	0.19	1.62	2.14
	Class Clustered	89.63	3.86	90.63	3.78	2.09	0.01	9.43	3.0
	NLL	90.15	3.55	89.84	3.38	0.90	0.18	11.43	1.93
APS	No Grouping	89.82	3.53	90.79	3.46	2.28	0.17	10.57	2.29
	Mondrian	89.96	3.47	91.04	3.41	2.78	0.16	1.83	2.14
	Class Clustered	89.16	4.38	90.19	4.29	2.41	0.01	9.07	3.0
	NLL	90.99	3.58	90.92	3.46	1.58	0.17	10.56	2.57
RAPS	No Grouping	90.66	4.03	91.00	4.04	2.93	0.00	14.0	2.86
	Mondrian	90.71	3.76	91.54	3.75	2.39	0.00	1.93	2.36
	Class Clustered	89.96	4.58	91.12	4.55	3.0	0.00	13.7	3.07
	NLL	90.66	3.87	90.02	3.72	1.49	0.00	12.54	1.71
SAPS	No Grouping	90.29	4.10	90.85	4.06	2.76	0.00	12.9	2.86
	Mondrian	91.04	4.2	92.01	4.12	3.01	0.00	1.77	3.14
	Class Clustered	89.82	5.42	90.18	5.33	1.16	0.00	8.22	2.29
	NLL	90.24	4.02	89.80	3.85	1.25	0.08	11.74	1.71

Table 10. Conformal Prediction (CP) results for Resnet50, trained on $\{P \cup A\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup A \cup C\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood using a GMM.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.47	2.25	90.30	2.21	4.01	0.45	8.65	2.43
	Mondrian	91.38	1.80	91.87	1.81	3.02	0.49	2.08	1.57
	Class Clustered	91.9	2.73	92.44	2.73	3.6	0.05	3.28	3.43
	NLL	92.09	2.4	91.72	2.38	1.72	0.46	7.16	2.57
APS	No Grouping	90.86	2.32	90.55	2.29	4.31	0.45	8.32	2.36
	Mondrian	92.16	2.11	92.47	2.09	2.72	0.46	2.36	1.71
	Class Clustered	91.06	2.78	91.50	2.78	2.10	0.06	4.21	2.71
	NLL	92.22	2.38	92.57	2.4	3.17	0.45	6.3	3.21
RAPS	No Grouping	88.85	2.73	89.06	2.72	5.86	0.0	9.63	3.5
	Mondrian	89.37	2.14	89.96	2.18	3.66	0.18	1.14	1.29
	Class Clustered	91.12	3.1	91.62	3.13	3.69	0.0	4.42	3.36
	NLL	89.57	2.71	89.42	2.63	2.33	0.09	9.0	1.86
SAPS	No Grouping	90.28	2.86	90.12	2.83	4.77	0.0	8.65	3.43
	Mondrian	90.86	2.36	90.84	2.72	2.72	0.22	1.44	2.14
	Class Clustered	90.21	2.39	90.43	2.39	3.05	0.20	1.43	2.00
	NLL	90.54	2.85	90.11	2.74	2.54	0.09	8.44	2.43

Table 11. Conformal Prediction (CP) results for Resnet50, trained on $\{P \cup A\}$ from PACS, and calibrated/tested on equal samples from $\{P \cup A \cup S\}$. Results are reported for different scoring functions and grouping methods. We set the number of groups to 6 for NLL grouping. The best and second best values are colored in green and red, respectively. We compute likelihood using a GMM.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.32	2.61	91.99	2.61	4.02	0.29	8.53	3.43
	Mondrian	90.49	2.23	91.63	2.24	2.73	0.36	1.68	1.86
	Class Clustered	90.79	2.21	92.12	2.22	3.33	0.35	4.63	2.43
	NLL	89.76	2.55	89.53	2.38	1.10	0.31	8.9	2.29
APS	No Grouping	90.24	2.62	91.91	2.63	4.01	0.29	8.01	3.5
	Mondrian	89.51	2.52	90.81	2.53	3.09	0.32	1.92	2.14
	Class Clustered	90.71	2.49	91.78	2.49	2.69	0.29	5.01	2.5
	NLL	89.72	2.58	89.61	2.44	1.02	0.30	7.86	1.86
RAPS	No Grouping	89.72	3.12	90.83	3.12	3.32	0.0	11.81	3.57
	Mondrian	90.11	2.61	91.05	2.65	2.46	0.12	2.27	1.57
	Class Clustered	90.45	2.80	91.36	2.83	2.53	0.02	3.45	3.0
	NLL	89.98	2.96	89.38	2.71	1.58	0.09	10.52	1.86
SAPS	No Grouping	90.06	3.08	91.40	3.07	3.28	0.0	9.88	2.57
	Mondrian	91.35	3.21	92.11	3.2	2.47	0.09	2.03	2.86
	Class Clustered	89.72	3.17	90.80	3.16	2.62	0.08	2.41	2.29
	NLL	88.78	2.93	88.17	2.62	2.72	0.15	10.49	2.29

C. Ablation Analysis

In this section, we examine the effect of the hyperparameter M , specifically the number of groups, on the likelihood-aware conformal prediction method. We present results on the ImageNet-C and ImageNet- $\bar{C}$ datasets using both ViT-B-16 and Swin-B architectures, varying M between 4, 5, and 6. The experimental setup follows the details outlined in Section 4.3 of the main paper. We note that, from the results presented below our observations made in Section 4.3 continue to hold regardless of the choice of M .

Table 12. CP results for SWINv2-B, trained on ImageNet and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C, are reported. We present results averaged across all corruption types and severity levels for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods with the number of groups set to 6 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	90.08	10.89	92.96	7.89	7.27	0.36	6.54	2.29
	Mondrian	90.85	21.69	93.13	17.35	6.14	0.12	5.07	3.0
	Class Clustered	89.91	13.57	92.64	10.04	6.68	0.3	5.24	2.29
	NLL	91.03	24.67	91.63	15.33	1.73	0.48	6.13	2.43
APS	No Grouping	90.17	47.41	91.06	35.32	2.72	0.40	5.11	2.29
	Mondrian	91.0	73.43	91.25	62.5	2.07	0.03	4.96	3.71
	Class Clustered	89.70	51.32	90.26	40.52	2.01	0.21	4.59	2.00
	NLL	90.35	56.01	90.42	36.97	0.54	0.42	4.78	2.00
RAPS	No Grouping	90.15	18.67	93.33	18.45	8.08	0.08	6.65	2.57
	Mondrian	90.95	35.55	92.8	35.57	5.47	0.0	4.98	3.07
	Class Clustered	90.24	29.5	92.74	29.47	6.5	0.0	5.10	2.64
	NLL	90.79	28.32	91.24	17.71	1.35	0.40	5.81	1.71
SAPS	No Grouping	90.13	18.42	93.32	18.16	8.13	0.12	6.62	2.71
	Mondrian	90.95	35.23	92.73	35.2	5.36	0.0	4.96	3.07
	Class Clustered	90.09	29.18	92.59	29.12	6.41	0.0	5.11	2.50
	NLL	90.71	28.27	91.09	17.70	1.19	0.39	5.31	1.71

Table 13. Conformal Prediction (CP) results averaged across all corruption types and severity levels for Swin-B, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 5 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	90.08	10.79	90.09	10.77	7.56	0.36	6.55	2.07
	Mondrian	90.85	21.5	90.85	21.48	6.03	0.12	5.08	3.0
	Class Clustered	89.90	13.45	89.91	13.44	7.0	0.3	5.25	2.21
	NLL	90.61	24.69	90.61	24.63	0.81	0.49	6.29	2.71
APS	No Grouping	90.23	47.25	90.23	47.21	2.78	0.41	5.11	2.00
	Mondrian	90.99	72.46	90.99	72.43	2.13	0.03	4.92	3.57
	Class Clustered	89.51	49.89	89.51	49.85	2.14	0.22	4.61	2.43
	NLL	90.24	55.23	90.24	55.15	0.40	0.43	4.77	2.00
RAPS	No Grouping	90.15	18.49	90.17	18.49	8.43	0.09	6.65	2.00
	Mondrian	90.95	35.23	90.96	35.23	5.67	0.0	4.98	3.21
	Class Clustered	90.21	29.68	90.22	29.68	6.66	0.0	5.14	2.64
	NLL	90.41	28.21	90.41	28.15	0.60	0.41	5.75	2.14
SAPS	No Grouping	90.12	18.24	90.13	18.24	8.48	0.12	6.62	2.29
	Mondrian	90.95	34.91	90.96	34.91	5.61	0.0	4.96	3.21
	Class Clustered	90.07	29.22	90.08	29.22	6.81	0.0	5.17	2.36
	NLL	90.33	28.17	90.33	28.10	0.52	0.39	5.25	2.14

Table 14. Conformal Prediction (CP) results averaged across all corruption types and severity levels for Swin-B, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 4 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	90.08	10.79	90.09	10.78	7.5	0.36	6.55	2.07
	Mondrian	90.85	21.5	90.85	21.49	5.99	0.12	5.08	3.0
	Class Clustered	89.90	13.45	89.91	13.44	6.8	0.3	5.25	2.21
	NLL	90.49	24.31	90.49	24.26	0.63	0.49	6.35	2.71
APS	No Grouping	90.13	46.78	90.14	46.76	2.7	0.41	5.12	2.29
	Mondrian	90.97	72.34	90.97	72.32	2.04	0.03	4.92	3.57
	Class Clustered	90.04	52.49	90.05	52.47	2.16	0.18	4.53	1.86
	NLL	90.22	55.42	90.22	55.35	0.33	0.43	4.76	2.29
RAPS	No Grouping	90.15	18.49	90.16	18.49	8.33	0.09	6.65	2.00
	Mondrian	90.95	35.23	90.96	35.23	5.60	0.0	4.97	3.21
	Class Clustered	90.4	30.12	90.41	30.12	6.64	0.0	5.10	2.93
	NLL	90.33	28.00	90.33	27.94	0.47	0.40	5.75	1.86
SAPS	No Grouping	90.11	18.24	90.13	18.24	8.4	0.12	6.63	2.00
	Mondrian	90.94	34.91	90.95	34.91	5.52	0.0	4.96	3.21
	Class Clustered	90.20	30.01	90.21	30.01	6.63	0.0	5.13	2.64
	NLL	90.26	27.95	90.26	27.90	0.40	0.39	5.25	2.14

Table 15. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 6 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	89.96	9.32	92.41	7.27	6.0	0.33	6.46	2.29
	Mondrian	90.78	16.79	92.54	12.9	4.54	0.17	5.19	3.29
	Class Clustered	90.07	11.54	92.16	8.74	5.11	0.28	5.14	2.14
	NLL	91.36	15.54	92.12	10.83	2.27	0.39	5.81	2.29
APS	No Grouping	90.01	46.15	90.30	34.84	1.23	0.40	4.69	1.93
	Mondrian	90.94	65.68	90.86	53.84	1.19	0.06	5.01	3.86
	Class Clustered	90.2	51.1	90.31	39.89	1.04	0.21	4.41	2.57
	NLL	90.16	49.03	90.17	36.03	0.46	0.40	4.60	1.64
RAPS	No Grouping	90.07	15.97	92.8	15.74	6.71	0.05	6.51	2.57
	Mondrian	90.9	33.05	92.28	32.48	3.87	0.0	5.03	2.93
	Class Clustered	90.14	26.77	92.18	26.3	5.05	0.0	5.10	2.64
	NLL	91.01	21.55	91.58	15.16	1.71	0.29	5.62	1.86
SAPS	No Grouping	90.06	15.85	92.79	15.59	6.73	0.08	6.52	2.57
	Mondrian	90.91	32.82	92.25	32.2	3.83	0.0	5.01	3.0
	Class Clustered	90.08	26.96	92.07	26.45	4.94	0.0	5.09	2.64
	NLL	90.91	21.54	91.42	15.17	1.56	0.29	5.39	1.79

Table 16. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 5 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	89.96	9.32	89.95	9.33	5.98	0.33	6.46	2.00
	Mondrian	90.78	16.79	90.77	16.8	4.08	0.17	5.19	3.14
	Class Clustered	90.07	11.54	90.06	11.55	4.84	0.28	5.14	2.14
	NLL	90.87	15.45	90.87	15.47	1.02	0.41	5.96	2.71
APS	No Grouping	90.03	46.27	90.03	46.31	1.16	0.40	4.68	1.64
	Mondrian	90.98	65.42	90.98	65.46	1.18	0.06	4.99	4.0
	Class Clustered	89.58	48.33	89.58	48.37	0.97	0.26	4.54	2.29
	NLL	90.11	48.88	90.11	48.93	0.34	0.40	4.59	2.07
RAPS	No Grouping	90.06	15.97	90.05	15.97	6.67	0.05	6.52	2.00
	Mondrian	90.89	33.05	90.89	33.05	3.56	0.0	5.03	3.21
	Class Clustered	90.21	27.41	90.20	27.41	4.83	0.0	5.11	2.64
	NLL	90.56	21.41	90.56	21.43	0.72	0.31	5.59	2.14
SAPS	No Grouping	90.06	15.85	90.04	15.85	6.7	0.08	6.52	2.21
	Mondrian	90.9	32.82	90.89	32.82	3.52	0.0	5.03	3.21
	Class Clustered	90.05	26.86	90.04	26.87	4.92	0.0	5.15	2.43
	NLL	90.47	21.40	90.47	21.42	0.65	0.30	5.32	2.14

Table 17. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet-C. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 4 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size (↓)	Mean Group Cov (%)	Mean Group Int. Size (↓)	Group CovGap (↓)	Singleton Hit Ratio (↑)	Class CovGap (↓)	Rank (↓)
THR	No Grouping	89.96	9.32	89.95	9.33	5.84	0.33	6.46	2.00
	Mondrian	90.78	16.79	90.77	16.8	4.05	0.17	5.19	3.43
	Class Clustered	90.07	11.54	90.06	11.55	4.87	0.28	5.14	2.14
	NLL	90.66	15.29	90.66	15.31	0.83	0.41	5.99	2.43
APS	No Grouping	89.99	46.02	89.98	46.05	1.12	0.40	4.69	1.79
	Mondrian	90.89	65.23	90.89	65.27	1.09	0.06	5.03	3.86
	Class Clustered	89.94	50.16	89.94	50.2	0.89	0.26	4.53	2.29
	NLL	90.1	48.81	90.1	48.85	0.30	0.40	4.59	2.07
RAPS	No Grouping	90.07	15.97	90.05	15.97	6.54	0.05	6.51	2.00
	Mondrian	90.89	33.04	90.89	33.05	3.50	0.0	5.04	3.21
	Class Clustered	90.27	27.84	90.26	27.84	4.78	0.0	5.10	2.64
	NLL	90.41	21.19	90.41	21.21	0.57	0.31	5.6	2.14
SAPS	No Grouping	90.05	15.85	90.04	15.85	6.55	0.08	6.52	2.00
	Mondrian	90.91	32.82	90.9	32.82	3.47	0.0	5.01	3.21
	Class Clustered	90.13	27.11	90.12	27.11	4.81	0.0	5.12	2.64
	NLL	90.31	21.18	90.31	21.20	0.49	0.30	5.34	2.14

Table 18. CP results for SWINv2-B, trained on ImageNet and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$, are reported. We present results averaged across all corruption types and severity levels for various scoring functions and grouping methods, with the number of groups set to 6 for NLL grouping. The best and second best values are colored in green and red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.12	5.88	92.83	4.44	6.79	0.40	6.56	2.43
	Mondrian	90.79	14.55	93.04	12.1	5.92	0.14	5.23	3.14
	Class Clustered	89.91	7.63	92.55	5.73	6.36	0.33	5.28	2.14
	NLL	90.89	13.45	91.43	8.6	1.57	0.49	6.25	2.29
APS	No Grouping	90.19	31.67	90.61	24.03	2.23	0.39	5.07	2.43
	Mondrian	90.99	55.64	90.9	47.72	1.95	0.03	5.0	3.71
	Class Clustered	89.68	35.09	89.82	27.9	1.89	0.24	4.59	2.14
	NLL	90.31	35.84	90.33	23.86	0.49	0.41	4.89	1.71
RAPS	No Grouping	90.20	9.03	93.11	8.81	7.33	0.12	6.6	2.43
	Mondrian	90.91	20.78	92.43	20.66	4.73	0.0	5.15	2.93
	Class Clustered	90.31	17.79	92.44	17.68	5.66	0.0	5.17	2.79
	NLL	90.67	14.38	91.00	9.17	1.13	0.41	5.76	1.86
SAPS	No Grouping	90.16	8.84	93.1	8.58	7.37	0.16	6.56	2.57
	Mondrian	90.91	20.57	92.34	20.38	4.57	0.01	5.11	3.0
	Class Clustered	90.06	18.28	92.17	18.09	5.49	0.0	5.18	2.57
	NLL	90.53	14.35	90.80	9.18	0.93	0.39	5.3	1.86

Table 19. Conformal Prediction (CP) results averaged across all corruption types and severity levels for Swin-B, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 5 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.12	5.88	90.13	5.88	6.92	0.40	6.56	2.29
	Mondrian	90.79	14.55	90.79	14.55	5.77	0.14	5.23	3.29
	Class Clustered	89.91	7.63	89.92	7.64	6.57	0.33	5.28	2.00
	NLL	90.52	13.52	90.51	13.52	0.75	0.50	6.42	2.43
APS	No Grouping	90.24	31.78	90.24	31.78	2.36	0.39	5.05	2.29
	Mondrian	90.99	55.18	91.0	55.18	2.11	0.03	4.97	3.71
	Class Clustered	89.35	33.41	89.35	33.41	2.03	0.24	4.62	2.29
	NLL	90.23	35.68	90.23	35.66	0.40	0.42	4.87	1.71
RAPS	No Grouping	90.21	9.03	90.22	9.03	7.54	0.12	6.6	2.00
	Mondrian	90.9	20.78	90.91	20.78	5.00	0.0	5.13	3.21
	Class Clustered	90.35	19.56	90.36	19.56	5.81	0.0	5.20	2.93
	NLL	90.33	14.43	90.33	14.43	0.57	0.41	5.68	1.86
SAPS	No Grouping	90.16	8.85	90.17	8.84	7.6	0.16	6.56	2.29
	Mondrian	90.9	20.56	90.9	20.56	4.89	0.01	5.11	3.14
	Class Clustered	90.03	17.86	90.04	17.86	5.78	0.0	5.17	2.43
	NLL	90.25	14.42	90.25	14.43	0.50	0.40	5.23	2.14

Table 20. Conformal Prediction (CP) results averaged across all corruption types and severity levels for Swin-B, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 4 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	90.12	5.88	90.13	5.88	6.92	0.40	6.56	2.29
	Mondrian	90.79	14.55	90.79	14.55	5.73	0.14	5.23	3.29
	Class Clustered	89.91	7.63	89.92	7.63	6.45	0.33	5.28	2.00
	NLL	90.41	13.58	90.41	13.57	0.61	0.50	6.44	2.43
APS	No Grouping	90.15	31.41	90.15	31.41	2.28	0.39	5.06	2.29
	Mondrian	90.94	55.0	90.95	55.0	1.99	0.03	4.98	3.71
	Class Clustered	90.11	37.33	90.11	37.32	1.95	0.22	4.56	2.00
	NLL	90.19	35.79	90.19	35.76	0.34	0.42	4.88	2.00
RAPS	No Grouping	90.20	9.03	90.21	9.03	7.53	0.12	6.6	2.00
	Mondrian	90.91	20.79	90.92	20.79	5.02	0.0	5.14	3.21
	Class Clustered	90.26	18.66	90.27	18.66	5.84	0.0	5.20	2.64
	NLL	90.28	14.47	90.28	14.46	0.48	0.41	5.69	2.14
SAPS	No Grouping	90.16	8.84	90.18	8.84	7.58	0.16	6.57	2.00
	Mondrian	90.92	20.56	90.93	20.56	4.91	0.01	5.10	3.14
	Class Clustered	90.31	19.03	90.32	19.03	5.93	0.0	5.15	3.0
	NLL	90.20	14.45	90.19	14.44	0.39	0.40	5.24	1.86

Table 21. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 6 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.92	6.99	92.35	5.43	6.01	0.36	6.41	2.43
	Mondrian	90.75	14.34	92.56	11.02	4.67	0.18	5.24	3.29
	Class Clustered	90.06	8.77	92.16	6.63	5.25	0.3	5.17	1.93
	NLL	91.38	11.58	92.16	8.34	2.30	0.42	5.81	2.36
APS	No Grouping	90.01	34.80	90.1	25.77	1.15	0.40	4.66	1.93
	Mondrian	90.88	53.94	90.66	43.85	1.33	0.07	5.04	4.0
	Class Clustered	90.15	39.61	90.09	30.39	1.09	0.22	4.44	2.43
	NLL	90.09	36.78	90.07	26.82	0.44	0.40	4.61	1.64
RAPS	No Grouping	90.04	11.39	92.7	11.16	6.55	0.08	6.38	2.43
	Mondrian	90.86	25.41	92.14	24.82	3.68	0.0	5.08	2.86
	Class Clustered	90.29	21.1	92.18	20.6	4.79	0.0	5.08	2.71
	NLL	90.91	15.48	91.43	11.18	1.60	0.33	5.49	2.00
SAPS	No Grouping	90.02	11.27	92.7	11.00	6.59	0.11	6.38	2.43
	Mondrian	90.86	25.25	92.09	24.6	3.58	0.0	5.08	3.07
	Class Clustered	90.15	21.31	92.01	20.83	4.72	0.0	5.10	2.64
	NLL	90.75	15.47	91.20	11.18	1.38	0.32	5.25	1.86

Table 22. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 5 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.92	6.99	90.75	7.00	6.02	0.36	6.41	2.29
	Mondrian	90.75	14.34	90.75	14.35	4.24	0.18	5.24	3.14
	Class Clustered	90.06	8.77	90.05	8.78	4.97	0.3	5.17	1.86
	NLL	90.9	11.59	90.9	11.62	1.10	0.43	5.95	2.71
APS	No Grouping	89.99	34.89	89.99	34.92	1.02	0.39	4.65	1.57
	Mondrian	90.91	53.44	90.91	53.47	1.34	0.07	5.04	4.0
	Class Clustered	89.51	37.24	89.51	37.27	1.07	0.27	4.55	2.71
	NLL	90.06	36.75	90.06	36.80	0.33	0.40	4.61	1.71
RAPS	No Grouping	90.04	11.39	90.03	11.39	6.51	0.08	6.37	2.00
	Mondrian	90.85	25.41	90.85	25.42	3.45	0.0	5.10	3.36
	Class Clustered	90.24	20.87	90.24	20.87	4.66	0.0	5.07	2.5
	NLL	90.51	15.26	90.51	15.28	0.71	0.33	5.44	2.14
SAPS	No Grouping	90.04	11.27	90.03	11.27	6.56	0.11	6.38	2.00
	Mondrian	90.87	25.24	90.87	25.25	3.40	0.0	5.06	3.21
	Class Clustered	90.09	21.26	90.09	21.27	4.69	0.0	5.12	2.64
	NLL	90.38	15.27	90.39	15.30	0.59	0.32	5.19	2.14

Table 23. Conformal Prediction (CP) results averaged across all corruption types and severity levels for ViT-B-16, trained on ImageNet, and calibrated/tested on equal samples from ImageNet-Val and corrupted samples from ImageNet- $\bar{C}$. Results are reported for different scoring functions (THR, APS, RAPS, SAPS) and grouping methods (NA Standard, Class Label, Ding et al., and NLL). We set the number of groups to 4 for NLL grouping. The best and second best values are colored in Green and Red, respectively.

Score Fn	Grouped by	Marginal Cov. (%)	Marginal Int. Size ($\downarrow$)	Mean Group Cov (%)	Mean Group Int. Size ($\downarrow$)	Group CovGap ($\downarrow$)	Singleton Hit Ratio ($\uparrow$)	Class CovGap ($\downarrow$)	Rank ($\downarrow$)
THR	No Grouping	89.92	6.99	90.75	7.00	5.76	0.36	6.41	2.29
	Mondrian	90.75	14.34	90.75	14.35	4.13	0.18	5.24	3.43
	Class Clustered	90.06	8.77	90.05	8.78	4.9	0.3	5.17	1.86
	NLL	90.7	11.29	90.7	11.32	0.87	0.43	6.01	2.43
APS	No Grouping	89.94	34.63	89.94	34.66	0.99	0.40	4.65	1.93
	Mondrian	90.84	53.33	90.84	53.37	1.27	0.07	5.06	4.0
	Class Clustered	89.92	39.2	89.92	39.23	0.95	0.27	4.54	2.57
	NLL	90.04	36.62	90.04	36.66	0.31	0.40	4.63	1.50
RAPS	No Grouping	90.03	11.39	90.03	11.39	6.26	0.08	6.38	2.00
	Mondrian	90.84	25.41	90.83	25.41	3.27	0.0	5.08	3.21
	Class Clustered	90.33	21.79	90.33	21.79	4.47	0.0	5.09	2.64
	NLL	90.36	15.08	90.37	15.11	0.54	0.33	5.44	2.14
SAPS	No Grouping	90.04	11.26	90.03	11.26	6.29	0.11	6.38	2.00
	Mondrian	90.86	25.25	90.85	25.25	3.21	0.0	5.08	3.21
	Class Clustered	90.07	21.76	90.06	21.76	4.45	0.0	5.12	2.64
	NLL	90.25	15.07	90.25	15.11	0.44	0.32	5.21	2.14