1. Supplementary Section

In this section we mention interesting observations and the training details which were used to report the experiments in the main paper. In addition to this, we also discuss the training schedule and additional experiments (for example transferability experiment on CIFAR100) which we could not fit in the main paper due to space constraints.

1.1. Underreported Baselines

Methods	10%	20%	30%	40%	Model
CIFAR10					
QBC	74	82.5	-	-	DenseNet121
VAAL	61.35	68.17	72.26	75.99	VGG16
Coreset	60	68	71	74	VGG16
RSB(ours)	69.16	77.34	80.91	82.05	VGG16
RSB-SR(ours)	82.16	85.09	89.43	91.16	VGG16
LLAL	81	87	-	-	ResNet18
CoreGCN	80	85.5	-	-	ResNet18
TA-VAAL	81	87.5	-	-	ResNet18
RSB-SR(ours)	**84.69**	**88.45**	**89.98**	**92.29**	ResNet18

Table 1. Reported Random Baseline vs our RSB results. We denote RSB results with strong regularization by RSB-SR.

In this section we analyze our random baseline (RSB) results with the random baselines reported by published methods in AL literature. From Tab. 1, it is evident that our strongly-regularized settings along with hyper-parameters tuned using AutoML yields strong baseline.

1.2. Training Algorithm

Algorithm 1 AL Training Schedule

```
Input $AL_{iter}$, Budget size $k$ and Oracle, $\mathcal{A}$
Split $\mathcal{D} \rightarrow \{T_r, T_s, V\}$
Split $T_r \rightarrow \{L_0^0, U_0^0\}$
Train a base classifier, $\mathcal{B}$ using only $L_0^0$
$\phi = \mathcal{B}$
while i $\in \{0 \ldots AL_{iter}\}$ do
    sample $\{x_j\}_{j=1}^k \in U_0^i$ using $\Psi(L_0^i, U_0^i, \phi)$
    $\{x_j, y_j\}_{j=1}^k \leftarrow \{x_j, \mathcal{A}(x_j)\}_{j=1}^k$
    $L_0^i \leftarrow L_0^i \cup \{x_j, y_j\}_{j=1}^k$
    $U_0^i \leftarrow U_0^i \setminus \{x_j, y_j\}_{j=1}^k$
    $\phi$ ←Initialize randomly
    while convergence do
        Train $\phi$ using only $L_0^i$
    end while
end while
```

For all reported experiments in the main paper we followed the algorithm described in Algorithm 1

1.3. Auto-ML Hyper-parameters

Here we enlist our hyper-parameters tuned using AutoML. To implement AutoML we used optuna framework extensively in our codebase.

- Learning rate: log-scale in range $[10^{-5}, 10^{-2})$
- Weight Decay : log-scale in range $[10^{-8}, 10^{-3})$

- Batch Size: Categorical values from [8,16,32....1024]
- Optimizer: Categorical values from [SGD, ADAM]
- Number of Transformation in randaug (RA_N) : Categorical values from [1,2,3,....15]
- Magnitude of Transformation in randaug (RA_M) : Categorical values from [1,2,3,....8]

1.4. Transferability Experiment

We mainly used three different architectures for classifier model *i.e*. VGG16, ResNet18 (R18) and Wide ResNet-28-2 (WRN)[1]. The VGG network was used as a source model whereas other two networks are used for target models. The results for CIFAR100 are reported in Table 2 which are achieved when we replace all the relu activations with leaky relu (negative slope set to 0.2) following (Oliver *et al*., 2018). We found CIFAR100 results to be significantly better with leaky relu activation, however, the same change does not affect the performance of CIFAR10.

	Source Model			Target Model					
	VGG16			WRN-28-2			R18-SR		
Methods ↓	20%	30%	40%	20%	30%	40%	20%	30%	40%
Random	46.72	50.63	55.27	47.87	56.53	57.84	60.17	64.8	69.33
Coreset	**48.2**	49.5	**56.99**	**51.25**	**58.39**	**60.56**	58.76	65.40	69.12
VAAL	39.32	52.17	55.73	49.13	57.72	55.71	59.76	61.36	67.15
QBC	46.53	**53.16**	55.54	49.02	53.51	57.05	**61.06**	**66.92**	**69.83**

Table 2. Transferability experiment on CIFAR100 dataset where source model is VGG16. The reported numbers are test accuracies corresponding to the best trained on CIFAR100 dataset. For best model hyper-parameters we perform random search over 50 trials(so for 4 AL iterations; we train 200 models in total). For this experiment we replace all relu activations with leaky relu (negative slope set to 0.2).

1.5. Optimizer settings

Different AL studies have reported different optimizer choices in their experiments. In this light, we analyze the optimizer chosen by AutoML and we analyze it on CIFAR10. The results are present in Table 3 of supplementary section. Contrary to the previous works where the optimizer is fixed in advance, we found that both Adam and SGD can sometimes work better than the other.

Optimizers	20%	30%	40%
CIFAR10			
SGD	4	3	5
ADAM	10	11	9
CIFAR100			
SGD	10	10	9
ADAM	4	4	5

Table 3. Analyzing best optimizer chosen by AutoML during random search over 50 trials for all the AL methods (VGG16 classifier) on CIFAR 10. As we implement 7 AL methods in both standard and strongly-regularized settings; so at each AL iteration we have a total of 14 best optimizers chosen.

[1] All Model definitions in AL Toolkit has been provided as a supplementary material

1.6. Noisy Oracle Experiments

In conjunction to RSB baselines (presented in main paper), we report performance of AL methods under noisy labels in active sets. The results are reported in Tab. 4 where we make the following observations: **(i)** it is quite evident that strongly-regularized model improves performance even in label corruptions scenarios. **(ii)** No AL method consistently outperforms the simple RSB baseline. **(iii)** Strong-regularization help reduce the performance difference between RSB and best AL method at a particular data split.

	without strong-regularization				**with strong-regularization**			
Methods ↓	10%	20%	30%	40%	10%	20%	30%	40%
				Noise: 10%				
RSB	69.16	72.08	76.62	80.88	82.16	84.96	86.06	89.13
Coreset	69.16	75.97	80.07	82.78	82.16	82.99	88.14	90.31
DBAL	69.16	**76.98**	**80.5**	**84.4**	82.16	85.04	88.04	**90.44**
BALD	69.16	75.29	80.24	84.04	82.16	82.15	**88.24**	89.45
VAAL	69.16	73.85	77.35	79.82	82.16	**85.32**	86.57	89.53
QBC	69.16	75.64	77.87	80.53	82.16	85.25	87.39	88.68
UC	69.16	75.94	80.42	81.92	82.16	82.61	85.19	88.62
				Noise: 20%				
RSB	69.16	69.42	75.89	79.61	82.16	77.39	85.9	85.12
Coreset	69.16	71.13	76.44	**80.07**	82.16	80.05	88.05	88.32
DBAL	69.16	71.26	76.24	82.2	82.16	81.31	83.67	91.14
BALD	69.16	70.34	**77.18**	79.86	82.16	**85.26**	**88.52**	**91.21**
VAAL	69.16	70.13	74.94	76.42	82.16	82.39	82.66	88.31
QBC	69.16	71.18	76.52	77.78	82.16	83.17	84.68	85.62
UC	69.16	**71.53**	75.48	78.48	82.16	84.57	83.00	88.42

Table 4. Mean accuracy on noisy oracle experiments on CIFAR10 with (n=3) repeated trials where the best hyper-parameters were found using the random search over 50 trials. We note that the noise is added in active sets drawn by AL methods. The strong-regularization experiments involve SWA and RA techniques.

1.7. Overlap in the active set

For the interested readers we plot the overlap in CIFAR10 active set sampled in the first AL iteration. As we do five runs for a labeled set partition, we therefore report the average overlap in Figure 1.

Methods ↓	10%	15%	20%	25%
		without RA + SWA		
RSB	57.89 ±0.13	61.55 ±0.32	64.51 ±0.13	66.52 ±0.2
Coreset	57.89 ±0.13	62.36 ±0.19	**65.42 ±0.19**	**67.8 ±0.23**
VAAL	57.89 ±0.13	**62.87 ±1.18**	65.11 ±0.74	67.08 ±0.42
		with RA + SWA		
RSB	60.1 ±0.09	64.59 ±0.62	67.14 ±0.18	69.2 ±0.11
Coreset	60.1 ±0.09	**64.98 ±0.50**	**67.97 ±0.25**	**70.12 ±0.13**
VAAL	60.1 ±0.09	64.88 ±0.53	67.45 ±0.26	69.37 ±0.08

Table 5. Effect of RA and SWA on ImageNet where annotation budget is 5% of training data. Reported results are averaged over 3 runs.

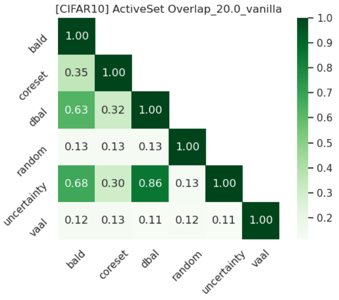

Figure 1. Overlap in CIFAR10 active set which is sampled during the first AL iteration.

1.8. Annotation Batch Size

Here we present the results for CIFAR10 and CIFAR100 in Table 6 for the experiment where annotation batch size is 5% relative to training data.

Methods ↓	15%	20%	25%	30%	35%	40%
	CIFAR10					
RSB	74.30 ± 0.88	78.27 ± 0.47	79.79 ± 0.64	81.86 ± 0.60	81.50 ± 0.45	83.21 ± 1.14
Coreset	74.56 ± 0.70	75.11 ± 0.92	**81.23** ± **0.27**	**82.58** ± **0.57**	83.9 ± 0.70	84.30 ± 0.56
DBAL	73.58 ± 0.81	**79.33** ± **0.61**	80.27 ± 1.10	81.78 ± 1.47	83.30 ± 0.75	83.86 ± 0.47
BALD	**75.43** ± **0.63**	79.19 ± 0.51	78.29 ± 0.63	81.69 ± 0.38	83.42 ± 1.54	**85.23** ± **0.41**
VAAL	74.07 ± 2.11	78.28 ± 1.00	78.88 ± 0.97	81.07 ± 0.61	80.98 ± 0.79	81.72 ± 2.33
QBC	72.63 ± 2.14	75.07 ± 2.07	76.95 ± 1.52	80.72 ± 0.34	81.76 ± 1.03	83.53 ± 0.59
UC	76.90 ± 1.12	78.14 ± 0.79	80.75 ± 0.62	81.47 ± 0.53	**84.60** ± **0.71**	83.13 ± 0.64
	CIFAR100					
RSB	35.15 ± 0.55	43.10 ± 0.46	49.33 ± 0.73	52.24 ± 0.56	51.76 ± 1.29	55.49 ± 0.64
Coreset	43.19 ± 0.65	42.58 ± 0.32	46.85 ± 0.83	**52.47** ± **0.58**	52.48 ± 0.93	57.45 ± 0.54
DBAL	35.83 ± 0.83	32.54 ± 1.92	42.93 ± 6.69	52.27 ± 3.59	54.58 ± 1.18	**57.68** ± **0.46**
BALD	37.55 ± 0.70	**43.86** ± **0.48**	49.79 ± 0.29	51.96 ± 0.81	54.75 ± 0.63	57.20 ± 0.90
VAAL	36.75 ± 1.36	37.05 ± 1.78	47.62 ± 1.07	47.20 ± 0.25	53.61 ± 0.44	52.87 ± 0.63
QBC	**38.91** ± **0.70**	43.57 ± 0.62	47.76 ± 0.61	51.16 ± 0.49	54.06 ± 0.33	56.51 ± 0.42
UC	36.52 ± 0.55	41.23 ± 0.89	**50.59** ± **0.50**	51.42 ± 0.42	**55.14** ± **0.97**	53.15 ± 0.36

Table 6. Mean Accuracy and Standard Deviation on CIFAR10/100 test set with annotation size as 5% of training set. Results reported are averaged over 5 runs where hyper-parameters are tuned in the first run using AutoML random search over 50 trials.

1.9. Unexplained performance degradation

In this section we discuss an counter-intuitive observation seen during AL iterations *i.e*. even with the increase in the labeled data, we sometimes observed the model performance (classification accuracy) degrading. More importantly, this observation was seen across different AL methods and datasets. For example on CIFAR10 from 20% to 30% AL cycle, the uncertainty method degrades its performance by **0.54%** (refer Tab. 8). Similarly on CIFAR100 and CIFAR10 from 30% to 35% AL cycle, the coreset and vaal method degrades its performance by **0.01%** and **0.09%** respectively (refer Tab. 6). Infact during our initial experiments without AutoML and strong-regularization, we observed such behaviour more frequent along-with high variance in accuracy and inconsistent ordinal ranking (by accuracy) across fractions of the data. These observations led us to employ AutoML and strong regularization, which helped reduce variance. We hypothesize that the performance drop could occur through a suboptimal active set selection by the AL method, as we do not interfere with active sets or settings used for AutoML (best of 50 experiments).

2. Additional Results

In the last we present the exact accuracies which were used to plot the Figure 1 in main paper. Tab. 7 to Tab. 11 reports the test accuracies for CIFAR10 dataset and Tab. 12 to Tab. 16 reports the test accuracies for CIFAR100 dataset.

Methods	20%	30%	40%
RSB	77.65 ± 0.82	81.39 ± 0.59	82.19 ± 1.55
Coreset	77.19 ± 1.93	**82.58 ± 0.67**	83.86 ± 1.08
DBAL	**78.81 ± 1.28**	80.99 ± 2.25	83.96 ± 2.01
BALD	78.35 ± 1.98	79.95 ± 1.43	84.29 ± 0.25
VAAL	75.89 ± 2.41	80.37 ± 0.34	81.75 ± 0.87
QBC	78.10 ± 0.73	80.31 ± 1.83	84.35 ± 0.64
UC	73.35 ± 4.84	81.98 ± 0.93	**84.49 ± 1.18**

Table 7. CIFAR10 Test Accuracy on L_0^0. The base model accuracy is 69.16.

Methods	20%	30%	40%
RSB	77.02 ± 0.71	80.50 ± 0.30	83.82 ± 0.37
Coreset	74.67 ± 0.82	81.14 ± 0.92	81.58 ± 1.19
DBAL	75.9 ± 0.25	80.58 ± 3.16	83.75 ± 0.88
BALD	76.19 ± 0.86	**83.26 ± 0.36**	**85.39 ± 0.97**
VAAL	76.88 ± 0.96	81.30 ± 0.29	82.63 ± 0.55
QBC	**78.38 ± 0.79**	81.39 ± 3.3	85.16 ± 0.77
UC	78.16 ± 0.85	81.80 ± 0.45	84.91 ± 0.69

Table 9. CIFAR10 Test Accuracy on L_2^0. The base model accuracy is 70.34.

Methods	20%	30%	40%
RSB	78.59 ± 0.91	81.81 ± 0.71	83.46 ± 0.18
Coreset	77.17 ± 1.82	81.37 ± 0.41	83.13 ± 1.54
DBAL	75.87 ± 0.61	83.00 ± 0.79	85.13 ± 1.25
BALD	78.49 ± 0.46	83.21 ± 0.66	85.06 ± 0.60
VAAL	73.67 ± 1.47	79.49 ± 1.27	82.98 ± 0.78
QBC	**78.61 ± 1.65**	**83.81 ± 0.49**	85.35 ± 0.82
UC	77.38 ± 1.17	81.82 ± 1.86	**85.62 ± 0.30**

Table 11. CIFAR10 Test Accuracy on L_4^0. The base model accuracy is 67.19.

Methods	20%	30%	40%
RSB	45.58 ± 0.19	**53.45 ± 0.28**	56.98 ± 0.31
Coreset	**46.05 ± 0.46**	52.04 ± 0.23	58.11 ± 0.12
DBAL	41.32 ± 0.23	52.16 ± 0.81	58.00 ± 0.68
BALD	43.57 ± 0.80	53.27 ± 0.12	56.87 ± 0.73
VAAL	42.70 ± 0.75	48.86 ± 1.61	54.81 ± 1.23
QBC	45.61 ± 0.74	53.31 ± 0.91	**58.21 ± 0.22**
UC	37.48 ± 0.45	53.01 ± 0.16	57.80 ± 0.09

Table 13. CIFAR100 Test Accuracy on L_1^0.The base model accuracy is 32.73.

Methods	20%	30%	40%
RSB	77.85 ± 0.65	81.68 ± 0.39	82.71 ± 0.42
Coreset	77.70 ± 1.31	82.78 ± 0.90	83.79 ± 0.74
DBAL	79.28 ± 0.78	81.16 ± 0.83	**85.58 ± 0.19**
BALD	78.67 ± 0.39	**82.95 ± 0.43**	84.11 ± 0.30
VAAL	76.50 ± 0.70	79.12 ± 0.62	82.86 ± 0.69
QBC	78.22 ± 1.84	82.68 ± 0.54	85.34 ± 1.26
UC	**80.03 ± 0.27**	79.49 ± 0.37	85.45 ± 0.69

Table 8. CIFAR10 Test Accuracy on L_1^0.The base model accuracy is 68.02.

Methods	20%	30%	40%
RSB	75.84 ± 1.91	80.93 ± 1.20	83.17 ± 0.52
Coreset	79.42 ± 0.47	81.62 ± 0.86	83.82 ± 0.18
DBAL	**79.48 ± 0.35**	82.27 ± 1.23	84.74 ± 0.14
BALD	77.58 ± 0.88	82.11 ± 0.65	84.58 ± 0.42
VAAL	77.45 ± 1.21	79.38 ± 1.08	82.90 ± 0.94
QBC	78.60 ± 0.43	**82.76 ± 0.92**	**85.54 ± 0.69**
UC	76.97 ± 0.79	81.35 ± 0.82	84.65 ± 0.30

Table 10. CIFAR10 Test Accuracy on L_3^0.The base model accuracy is 68.19.

Methods	20%	30%	40%
RSB	46.67 ± 0.30	51.43 ± 0.81	55.06 ± 0.35
Coreset	**47.33 ± 0.64**	49.73 ± 0.92	57.05 ± 0.40
DBAL	45.53 ± 2.33	51.04 ± 0.49	**58.06 ± 0.51**
BALD	47.10 ± 1.24	50.40 ± 0.88	55.65 ± 0.34
VAAL	39.73 ± 0.43	50.95 ± 0.88	55.23 ± 0.63
QBC	46.04 ± 0.57	**53.20 ± 0.38**	57.63 ± 0.49
UC	41.37 ± 1.29	52.97 ± 0.83	55.45 ± 0.62

Table 12. CIFAR100 Test Accuracy on L_0^0. The base model accuracy is 34.73.

Methods	20%	30%	40%
RSB	44.71 ± 0.64	50.01 ± 0.36	56.27 ± 0.84
Coreset	46.00 ± 0.79	53.48 ± 0.61	57.22 ± 0.69
DBAL	44.06 ± 0.39	49.29 ± 1.00	57.40 ± 0.34
BALD	**46.78 ± 0.52**	52.34 ± 0.90	54.97 ± 0.98
VAAL	44.75 ± 0.57	49.72 ± 0.40	55.77 ± 0.62
QBC	46.20 ± 0.72	53.15 ± 0.90	**57.96 ± 0.65**
UC	43.94 ± 0.60	**53.75 ± 0.50**	55.10 ± 0.95

Table 14. CIFAR100 Test Accuracy on L_2^0. The base model accuracy is 34.66.

Methods	20%	30%	40%
RSB	42.46 ± 0.44	52.66 ± 0.66	54.15 ± 0.43
Coreset	45.98 ± 0.83	**54.34 ± 0.53**	56.96 ± 0.95
DBAL	45.49 ± 0.51	48.84 ± 0.42	57.65 ± 0.46
BALD	**47.21 ± 1.26**	52.53 ± 0.42	55.39 ± 0.72
VAAL	44.93 ± 1.61	46.27 ± 0.72	56.65 ± 0.60
QBC	46.50 ± 0.56	53.49 ± 0.53	**57.68 ± 0.51**
UC	46.96 ± 0.41	53.07 ± 0.57	56.35 ± 0.79

Table 15. CIFAR100 Test Accuracy on L_3^0.The base model accuracy is 30.44.

Methods	20%	30%	40%
RSB	41.15 ± 0.89	50.61 ± 0.40	56.77 ± 0.55
Coreset	45.72 ± 0.77	52.22 ± 0.54	56.28 ± 0.45
DBAL	44.71 ± 0.57	52.33 ± 0.49	56.52 ± 0.51
BALD	40.35 ± 0.75	51.87 ± 0.60	57.40 ± 0.40
VAAL	44.86 ± 1.69	51.32 ± 1.54	53.82 ± 1.08
QBC	**45.93 ± 0.46**	**53.12 ± 0.55**	**57.78 ± 0.49**
UC	43.07 ± 0.74	49.89 ± 0.79	56.15 ± 0.52

Table 16. CIFAR100 Test Accuracy on L_4^0.The base model accuracy is 34.85.