7 Supplementary Material

7.1 Experimental results: Q1

This Supplement contains additional details about our experimental results for Q1 (in Sec. 4.2). First, we analyze the learning curves of SLADe and the baselines on all 21 datasets. Second, we provide two additional tables counting the wins of SLADe against the baselines similar to Table 2, but for 0% and 10% noise.

Learning curves. Figure 3 illustrates the learning curves when no noise is present. We can conclude that SLADe performs visibly better than the baselines on one, and performs similar to the baselines on 14 datasets. On three out of the other six datasets, it is only outperformed by the GP, which is easily able to model the prediction probabilities due to the fact that no noise is present. However, this setting is highly unlikely. Additionally, on 14 of the datasets we achieve an AUROC score at 60% that is significantly higher than at 5% of labels, showing that SLADe learns from the soft labels that it gathers. Out of the other seven datasets, we see that on three datasets we already achieve an AUROC score close to 1 at 5% of labels, leaving no room for improvement.

Figure 4 illustrates the learning curves with 10% noise added to the labels. Results show that SLADe performs visibly better than the baselines on seven datasets, and performs similarly to the baselines on eight datasets. Comparing that to Figure 3, we now outperform the baselines on six more datasets, showing SLADe's noise-robustness. Additionally, on 13 of the datasets we achieve an AUROC score at 60% labels that is significantly higher than at 5% of labels, showing that SLADe learns from the soft labels that it gathers. This means that even adding a bit of noise does not hold SLADe from learning from the provided soft labels.

Figure 5 finally displays the learning curves with 20% noise added to the labels. Under this level of noise, SLADe performs much better than the baselines on seven datasets, and performs similar to the baselines on ten datasets. Additionally, on 15 of the datasets we achieve an AUROC score at 60% labels that is significantly higher than at 5% of labels, showing that SLADe learns from the soft labels that it gathers. The learning curves become flatter when noise is added. However, comparing SLADe to HIF that only learns on seven datasets, we can confidently claim that SLADe is more resistant to noise. Additionally, we notice that GP needs many labels to achieve a good performance under the influence of noise. In an Active Learning setting, this is not always possible as labels are usually extremely costly.

Wins/Draws/Losses. Table 6 and 7 show the wins, draws and losses of SLADe against the baselines in settings with respectively 10% and 0% noise. The results for 10% noise are similar to those for 20% noise that were covered in the main text. At any label percentage, SLADe outperforms (W) or performs similarly (D) to all the baselines on at least 15 datasets out of 21. At 60% of labels SLADE draws more often than it wins against GP, but stays the clear winner with eight wins for SLADE against four for GP. At 0% noise, we notice how SLADE is outperformed by GP, especially at high label percentages (≥ 40%). However, such a setting occurs rarely and SLADE still wins or performs similarly (D) on at least 12 datasets for high label percentages.

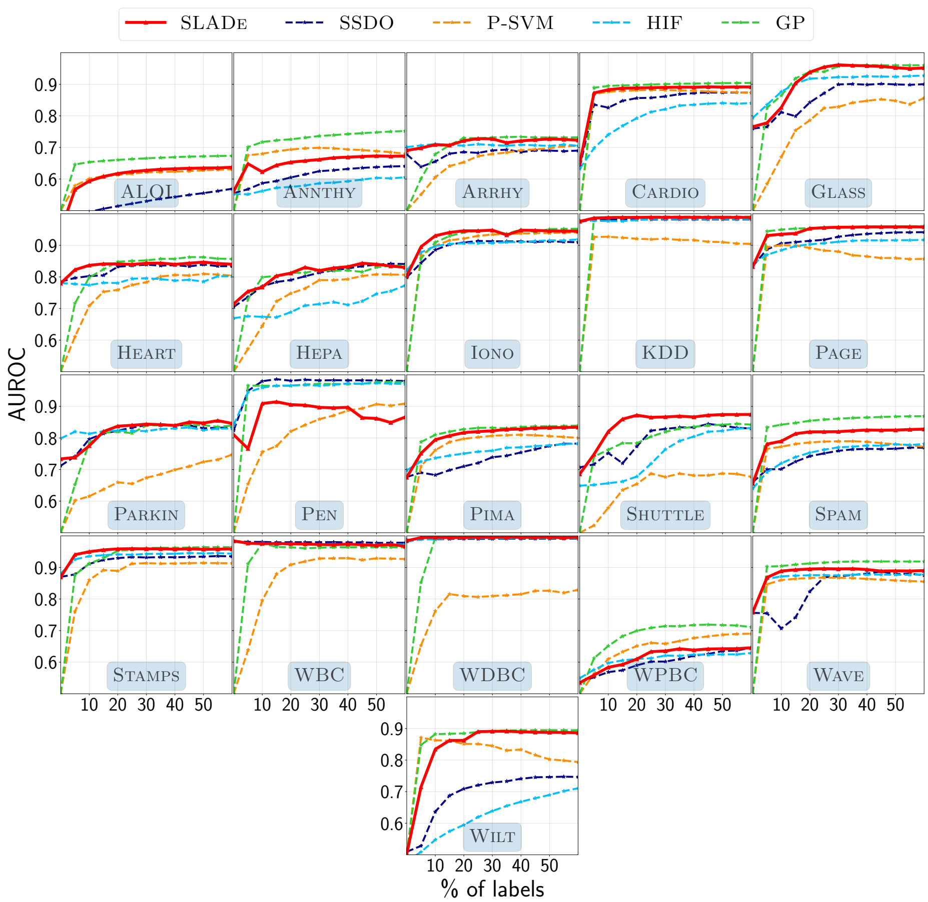

Fig. 3. Learning curves for all methods on all 21 datasets under no noise. On the x-axis we vary the label percentage, while on the y-axis we report the average AUROC (higher is better).

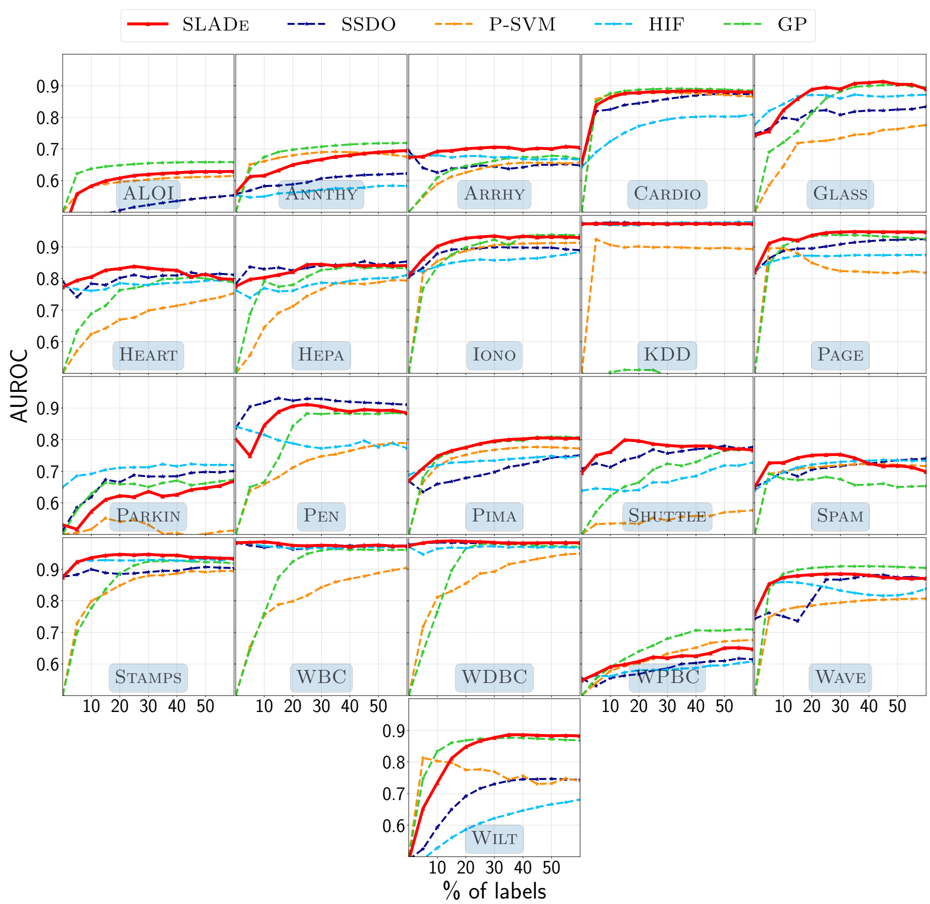

Fig. 4. Learning curves for all methods on all 21 datasets under 10% noise. On the x-axis we vary the label percentage, while on the y-axis we report the average AUROC (higher is better).

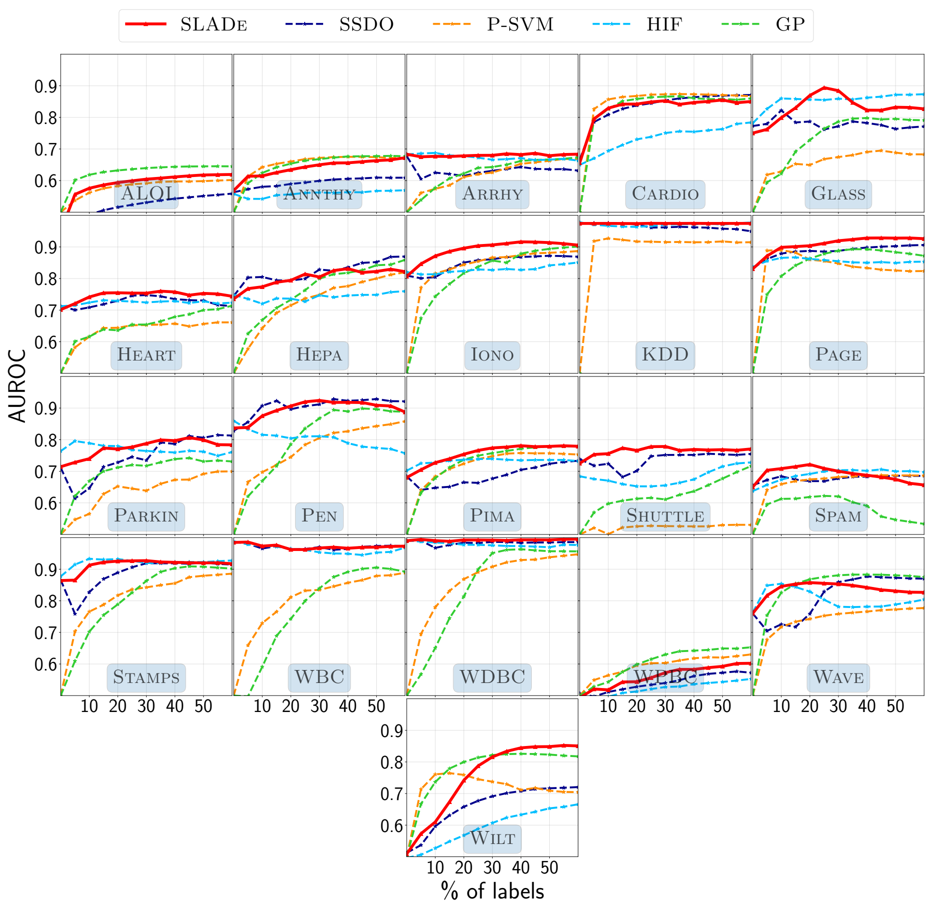

Fig. 5. Learning curves for all methods on all 21 datasets under 20% noise. On the x-axis we vary the label percentage, while on the y-axis we report the average AUROC (higher is better).

Table 6. Wins (W), Draws (D), and Losses (L) of **SLADe against each baseline** in terms of average AUROC per dataset, for each label percentage, under 10% of noise. A draw means that the absolute difference in AUROC is ≤ 0.01.

	SSDO			**P-SVM**			**HIF**			**GP**		
Labels	W	D	L	W	D	L	W	D	L	W	D	L
5%	**14**	4	3	**17**	1	3	**11**	7	3	**16**	0	5
10%	**16**	2	3	**17**	1	3	**17**	2	2	**13**	2	6
15%	**15**	3	3	**17**	3	1	**18**	2	1	**12**	3	6
20%	**16**	3	2	**18**	2	1	**18**	2	1	**12**	3	6
25%	**16**	3	2	**18**	2	1	**18**	2	1	**11**	5	5
30%	**15**	4	2	**18**	1	2	**18**	2	1	**10**	6	5
35%	**15**	4	2	**18**	1	2	**17**	3	1	**11**	5	5
40%	**13**	6	2	**17**	3	1	**17**	3	1	**9**	7	5
45%	**10**	6	5	**17**	3	1	**17**	2	2	**10**	6	5
50%	**10**	7	4	**18**	2	1	**16**	3	2	**10**	6	5
55%	**10**	7	4	**19**	1	1	**15**	4	2	**8**	8	5
60%	**10**	5	6	**19**	0	2	**15**	4	2	**8**	9	4

Table 7. Wins (W), Draws (D), and Losses (L) of **SLADe against each baseline** in terms of average AUROC per dataset, for each label percentage, under no noise (0%). A draw means that the absolute difference in AUROC is ≤ 0.01.

	SSDO			**P-SVM**			**HIF**			**GP**		
Labels	W	D	L	W	D	L	W	D	L	W	D	L
5%	**15**	5	1	**16**	2	3	**12**	5	4	**8**	2	11
10%	**15**	4	2	**16**	2	3	**14**	3	4	**5**	4	12
15%	**16**	4	1	**16**	3	2	**14**	5	2	**5**	6	10
20%	**16**	4	1	**17**	2	2	**16**	4	1	**3**	10	8
25%	**15**	5	1	**17**	2	2	**17**	3	1	**5**	7	9
30%	**14**	6	1	**17**	2	2	**17**	3	1	**2**	11	8
35%	**14**	6	1	**17**	2	2	**16**	4	1	**1**	11	9
40%	**14**	6	1	**17**	2	2	**16**	4	1	**2**	10	9
45%	**14**	6	1	**16**	2	3	**17**	3	1	**2**	11	8
50%	**14**	6	1	**16**	2	3	**17**	3	1	**1**	12	8
55%	**13**	7	1	**16**	2	3	**17**	3	1	**2**	10	9
60%	**14**	4	3	**16**	3	2	**17**	3	1	**1**	12	8