Supplementary Material

Breno Caetano da Silva
Jorge Luiz Franco
Hermes Manoel Galvão Castelo Branco
Eric K. Tokuda
Alexandre Delbem

July 16, 2024

Contents

1 Introduction **2**

2 ADABOOST **3**
2.1 seeds . . . 3
2.2 sonar . . . 3
2.3 parkinson . . . 3
2.4 wine . . . 4
2.5 haberman . . . 4
2.6 glass . . . 4
2.7 heart . . . 4
2.8 musk . . . 5

3 Linear Regression **5**
3.1 seeds . . . 5
3.2 sonar . . . 5
3.3 parkinson . . . 6
3.4 wine . . . 6
3.5 glass . . . 6
3.6 heart . . . 7
3.7 musk . . . 7
3.8 haberman . . . 7

4 Gaussian Process **8**
4.1 seeds . . . 8
4.2 sonar . . . 8
4.3 parkinson . . . 8
4.4 wine . . . 9
4.5 haberman . . . 9
4.6 glass . . . 9
4.7 heart . . . 10
4.8 musk . . . 10

5 SVM **10**
5.1 seeds . . . 10
5.2 sonar . . . 11
5.3 parkinson . . . 11
5.4 haberman . . . 11
5.5 glass . . . 12
5.6 heart . . . 12
5.7 wine . . . 12
5.8 musk . . . 13

6 Decision Tree 13
6.1 heart 13
6.2 musk 13

7 MLP 14
7.1 seeds 14
7.2 sonar 14
7.3 parkinson 14
7.4 wine 15
7.5 haberman 15
7.6 glass 15
7.7 heart 16
7.8 musk 16

8 Random Forest 16
8.1 seeds 16
8.2 parkinson 17
8.3 wine 17
8.4 glass 17
8.5 heart 18
8.6 musk 18

9 KNN 18
9.1 seeds 18
9.2 sonar 19
9.3 parkinson 19
9.4 wine 19
9.5 haberman 20
9.6 glass 20
9.7 heart 20
9.8 musk 21

1 Introduction

In Active Learning (AL), selecting an effective initial set of labeled samples is crucial for developing robust models with limited data. Traditional AL methods often rely on an oracle to label samples, but without a well-chosen initial set, model performance can be suboptimal.

This supplementary material presents PhyIL, our novel approach using phylogram analysis and community detection to enhance cold-start selection in AL. By choosing samples that represent data diversity and complexity, PhyIL ensures a robust initial model.

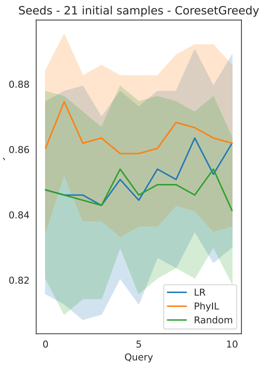
We evaluated PhyIL on nine diverse real-world datasets: seeds, sonar, parkinson, wine, haberman, glass, heart, and musk. These datasets were chosen for their variety in features and challenges, offering a comprehensive test of PhyIL's effectiveness. For example, the seeds dataset involves grain classification, while the sonar dataset deals with the differentiation between mines and rocks based on sonar signals. The parkinson dataset contains speech signals to detect Parkinson's disease, and the wine dataset is used for wine classification based on chemical properties.

We compared PhyIL against established methods: Query By Committee (QBC), which uses multiple models to select informative samples; QUIRE, which selects samples based on label ambiguity; Coreset-Greedy, which optimizes sample selection for model accuracy; and Uncertainty Sampling, which picks samples where the model is least certain.

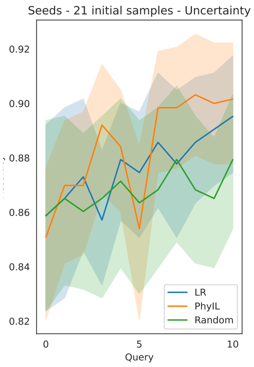
Our results show that PhyIL outperforms these methods, leading to better initial sample selection and improved model performance. This document provides detailed results for various classifiers, including ADABOOST, Linear Regression, Gaussian Process, Support Vector Machine (SVM), Decision Tree, Multi-Layer Perceptron (MLP), Random Forest, and K-Nearest Neighbors (KNN), demonstrating PhyIL's advantages across these datasets.

2 ADABOOST

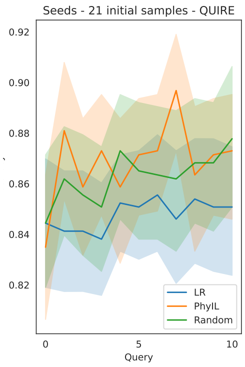
2.1 seeds



(a)

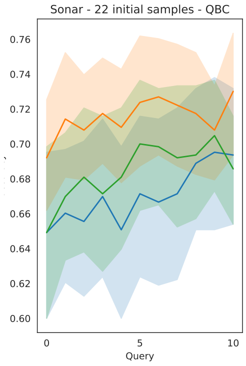


(b)

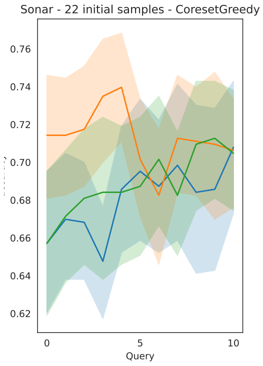


(c)

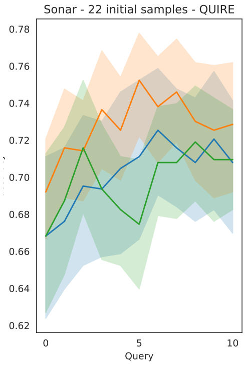
2.2 sonar



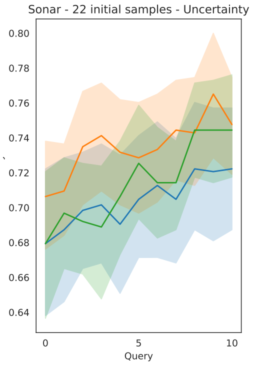
(a)



(b)

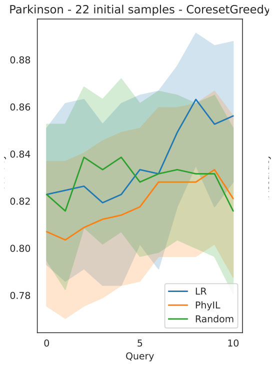


(c)

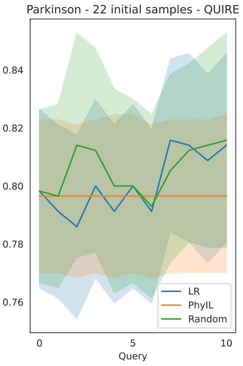


(d)

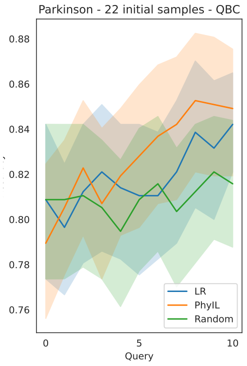
2.3 parkinson



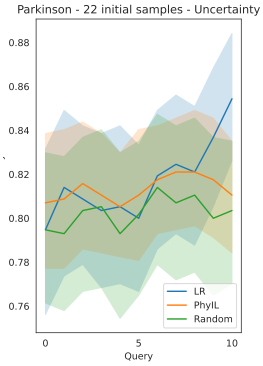
(a)



(b)

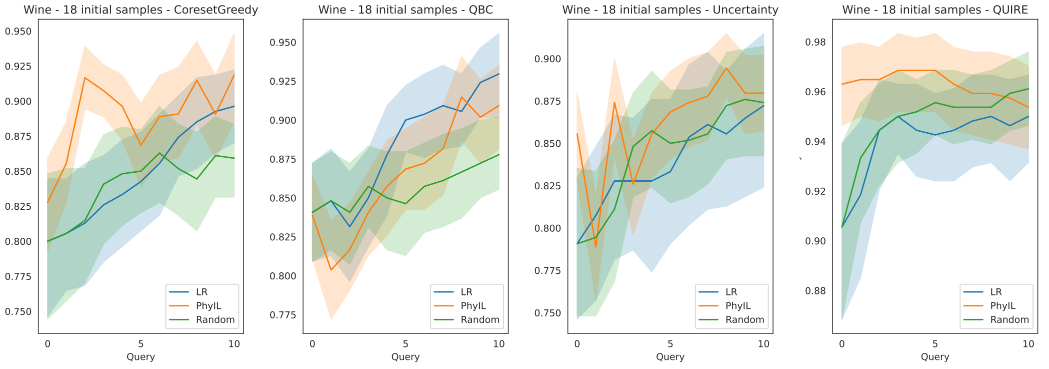


(c)



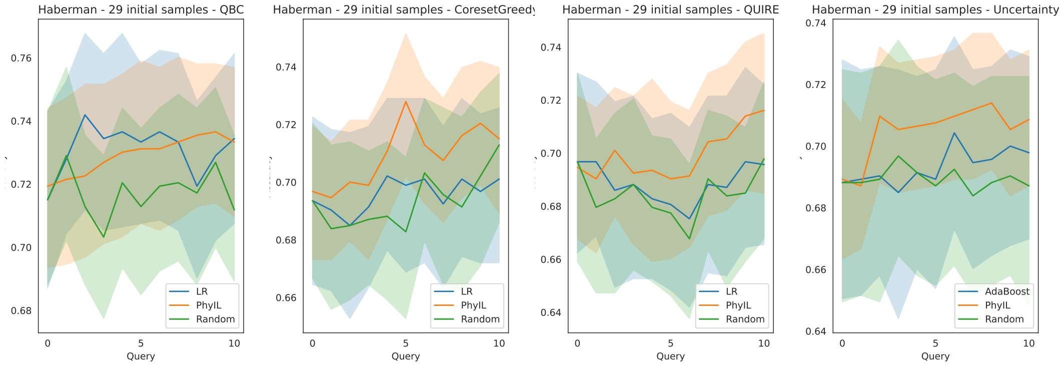
(d)

2.4 wine



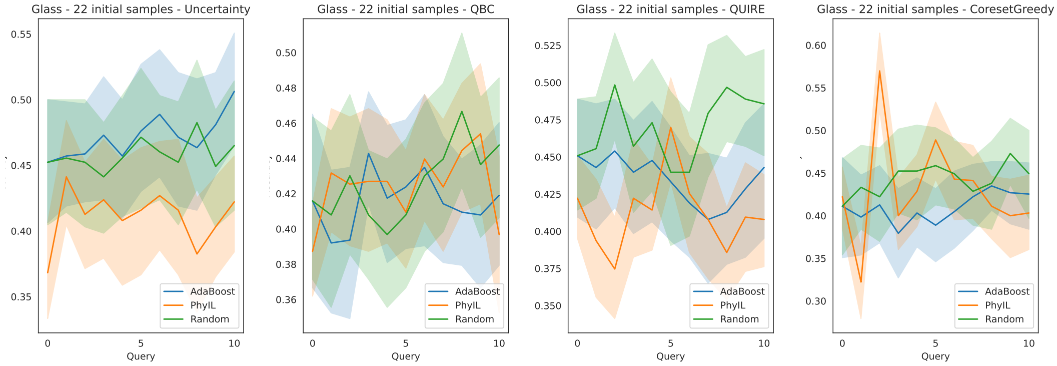
(a) (b) (c) (d)

2.5 haberman



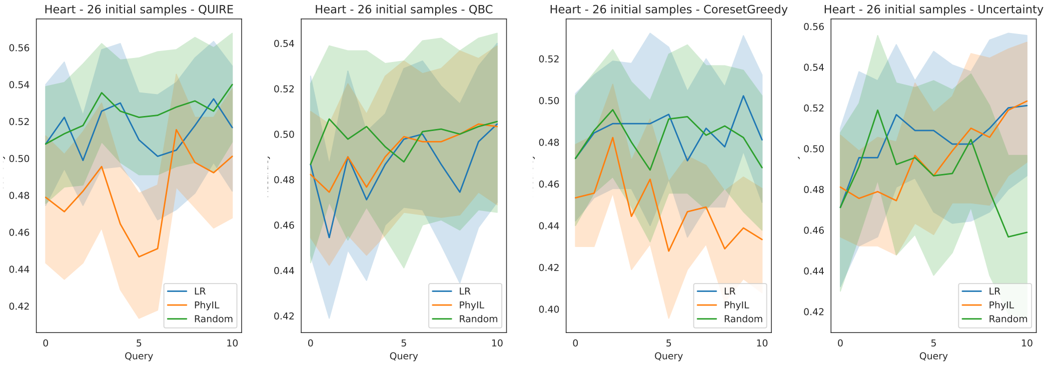
(a) (b) (c) (d)

2.6 glass



(a) (b) (c) (d)

2.7 heart



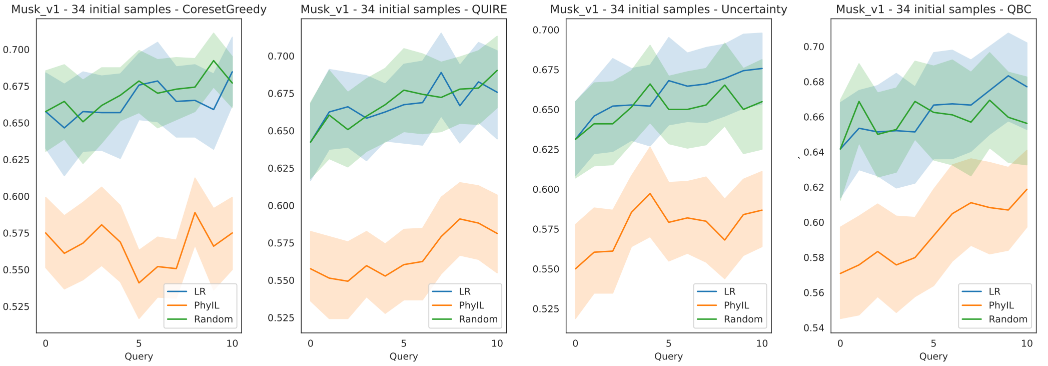
(a)

(b)

(c)

(d)

2.8 musk



(a)

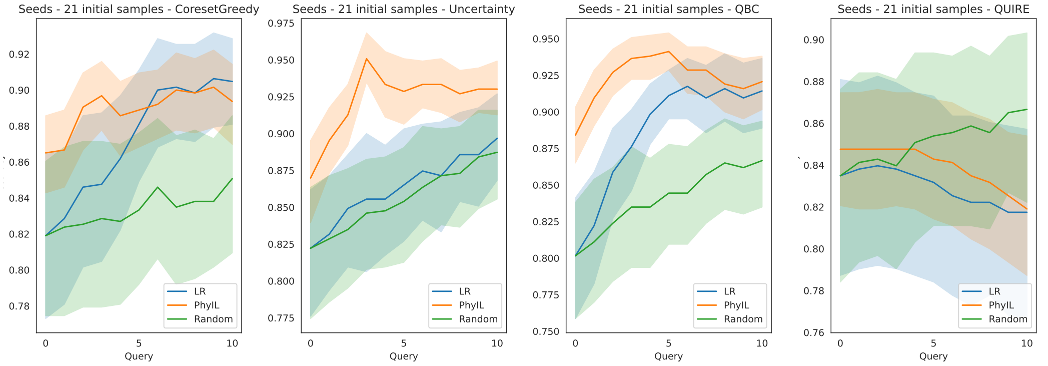
(b)

(c)

(d)

3 Linear Regression

3.1 seeds



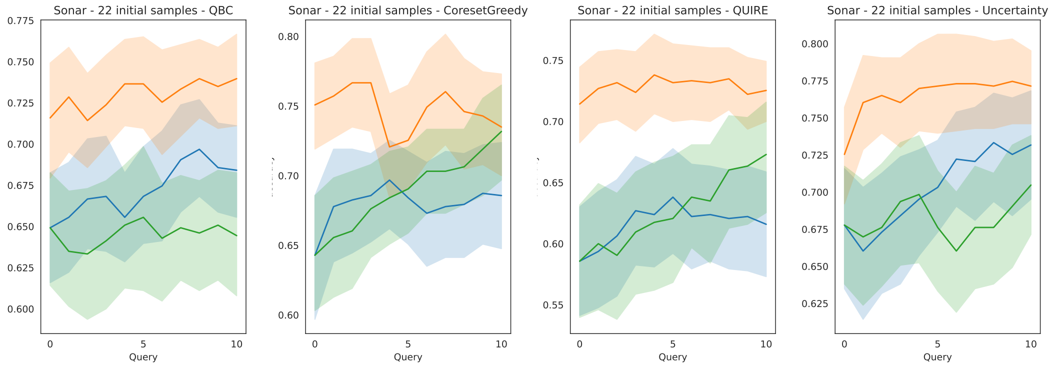
(a)

(b)

(c)

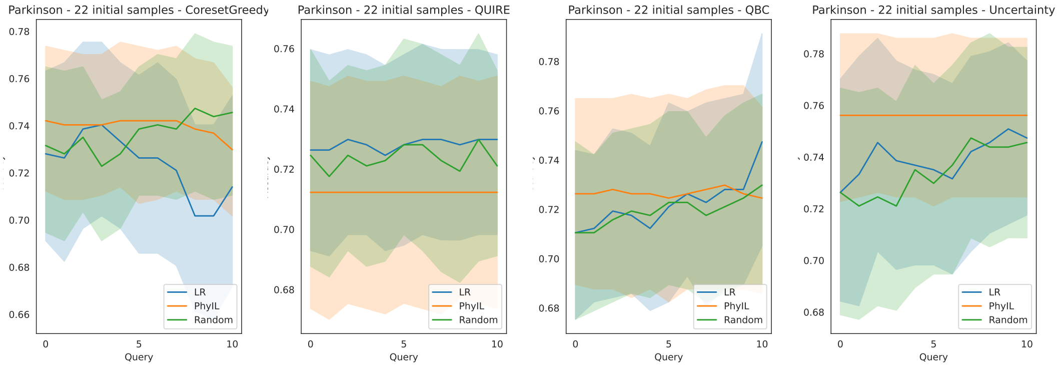
(d)

3.2 sonar



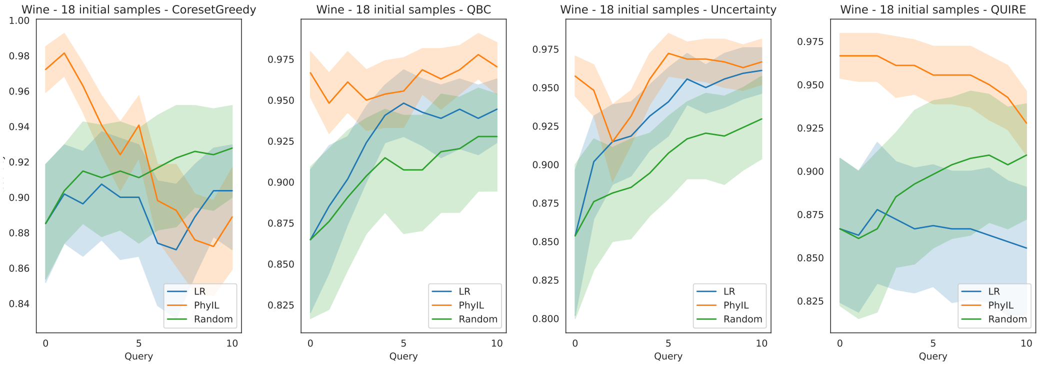
(a) (b) (c) (d)

3.3 parkinson



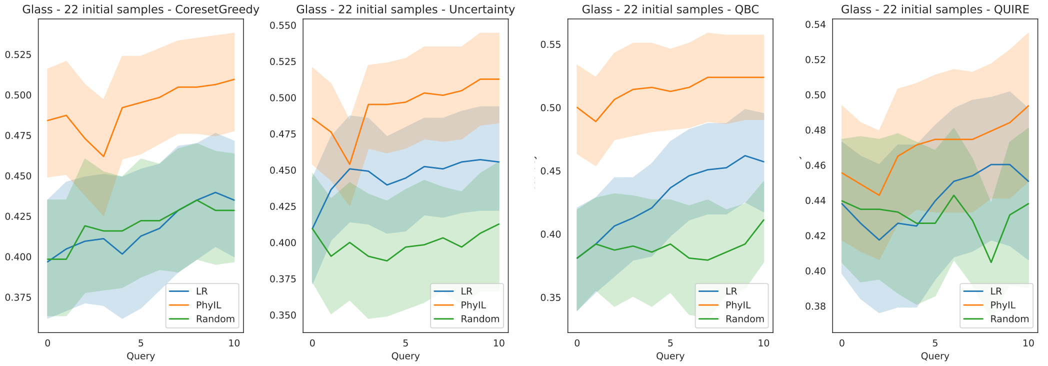
(a) (b) (c) (d)

3.4 wine



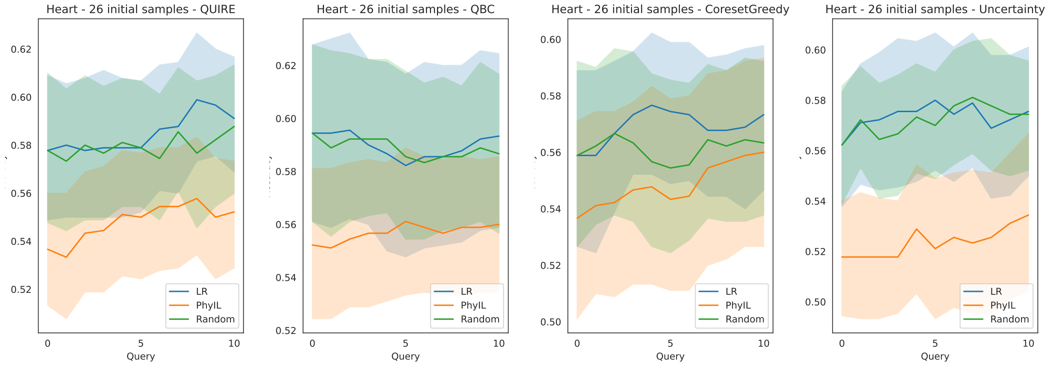
(a) (b) (c) (d)

3.5 glass



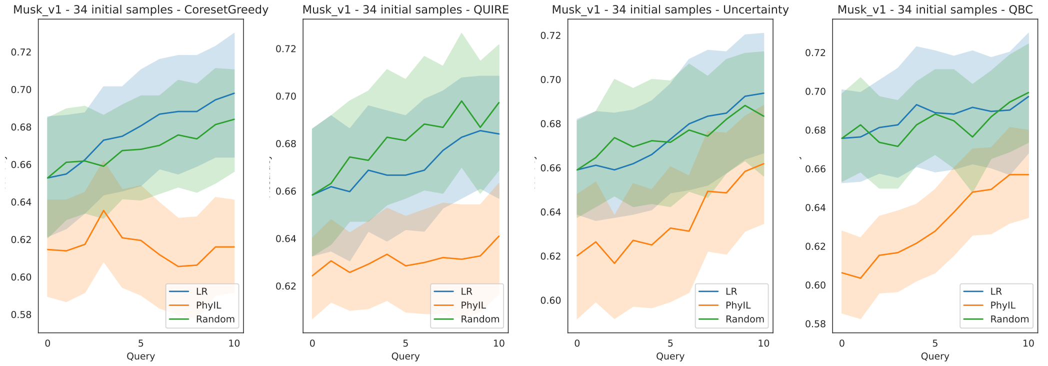
(a) (b) (c) (d)

3.6 heart



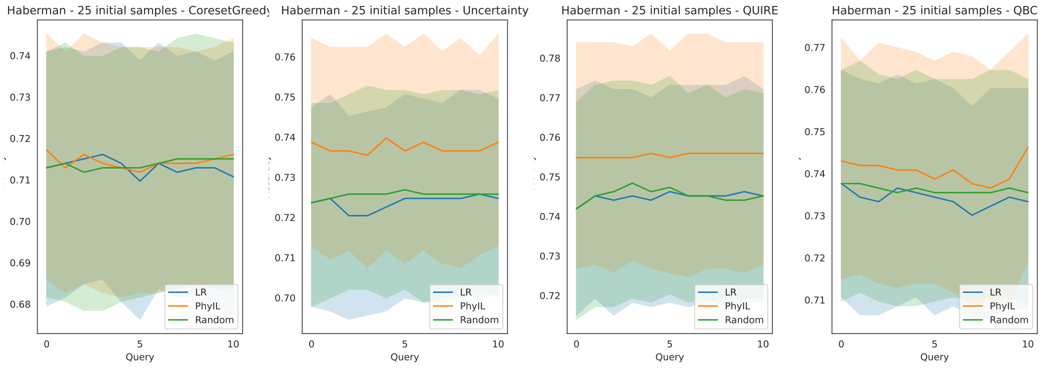
(a) (b) (c) (d)

3.7 musk



(a) (b) (c) (d)

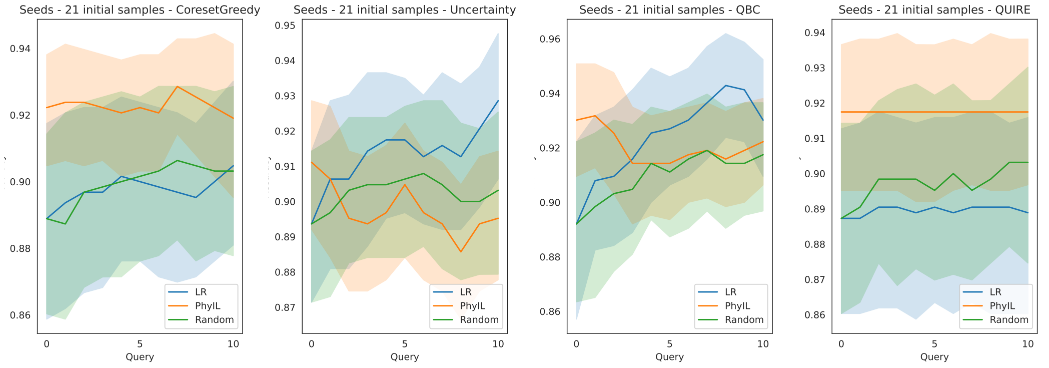
3.8 haberman



(a) (b) (c) (d)

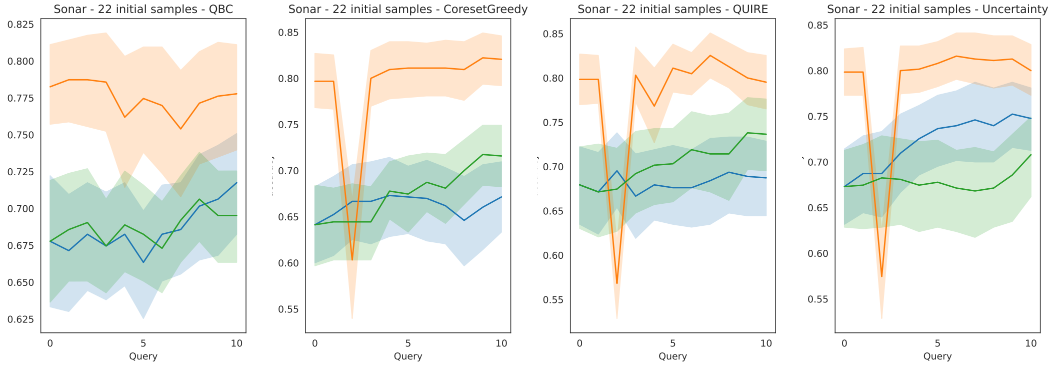
4 Gaussian Process

4.1 seeds



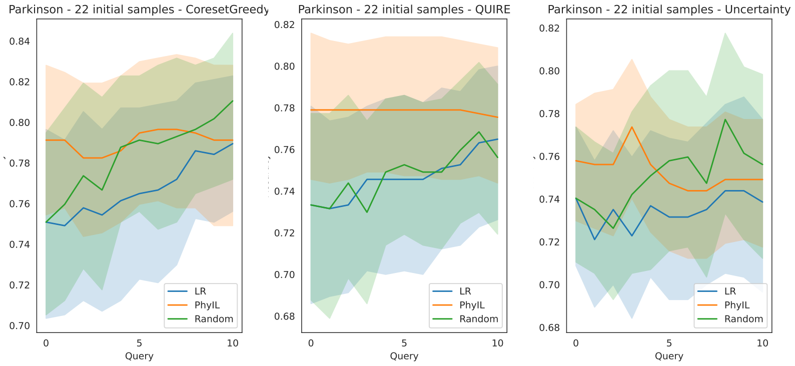
(a) (b) (c) (d)

4.2 sonar



(a) (b) (c) (d)

4.3 parkinson

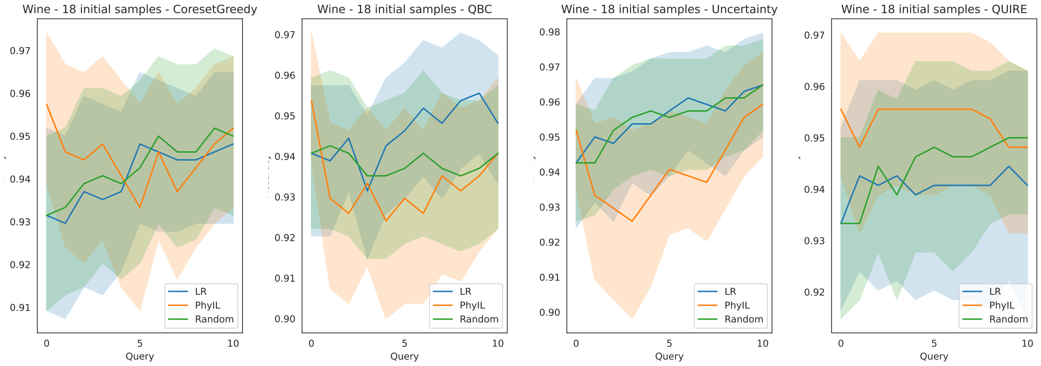

Parkinson - 22 initial samples - CoresetGreedy

Parkinson - 22 initial samples - QUIRE

Parkinson - 22 initial samples - Uncertainty

(a) (b) (c)

4.4 wine


Wine - 18 initial samples - CoresetGreedy

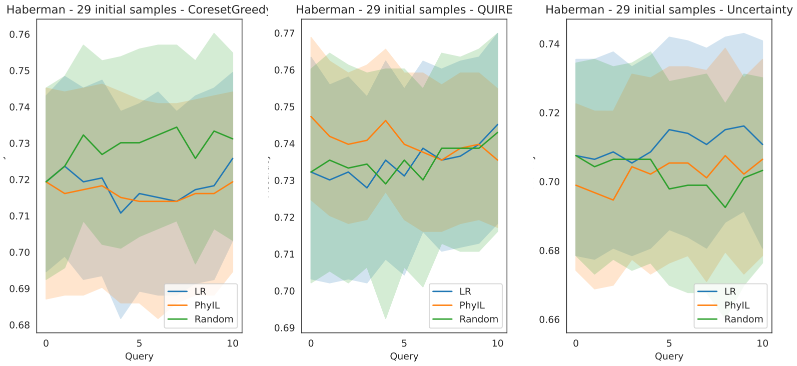
Wine - 18 initial samples - QBC

Wine - 18 initial samples - Uncertainty

Wine - 18 initial samples - QUIRE

(a) (b) (c) (d)

4.5 haberman

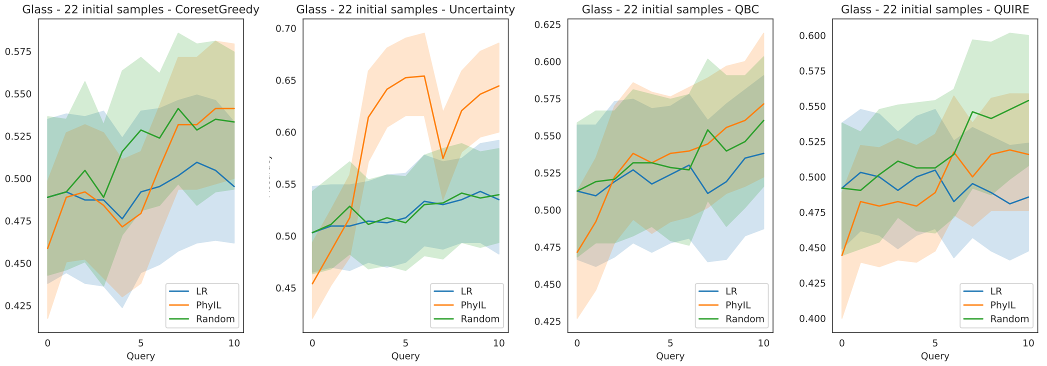

Haberman - 29 initial samples - CoresetGreedy

Haberman - 29 initial samples - QUIRE

Haberman - 29 initial samples - Uncertainty

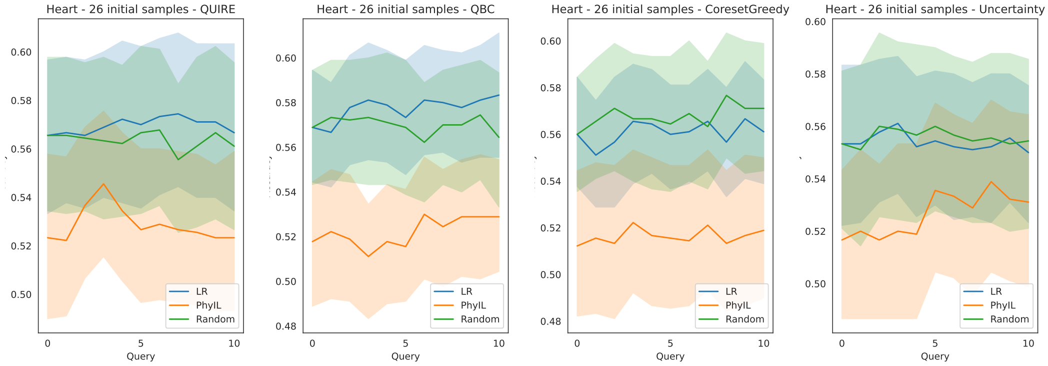
(a) (b) (c)

4.6 glass



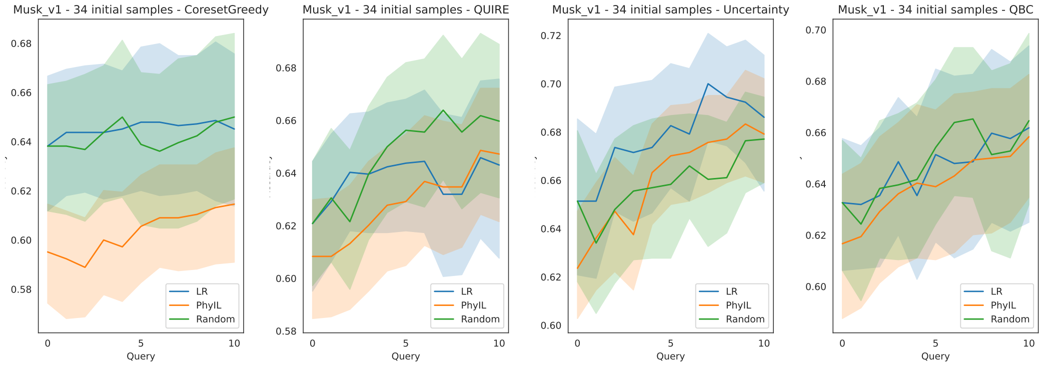
(a) (b) (c) (d)

4.7 heart



(a) (b) (c) (d)

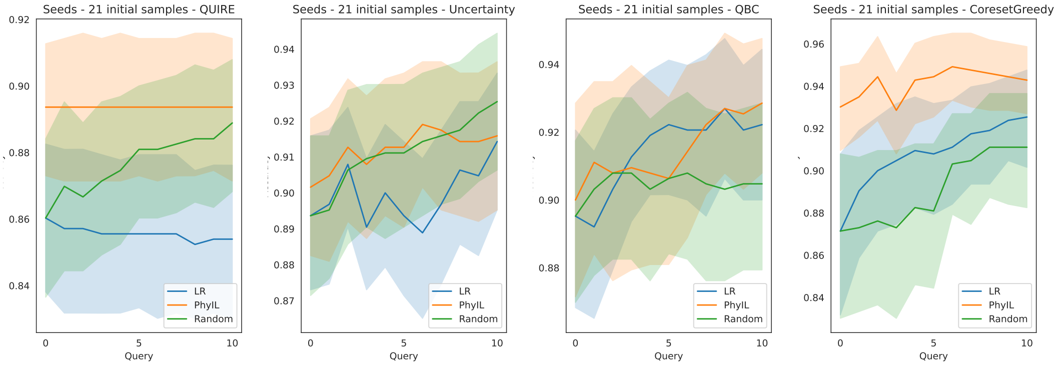
4.8 musk



(a) (b) (c) (d)

5 SVM

5.1 seeds


Seeds - 21 initial samples - QUIRE

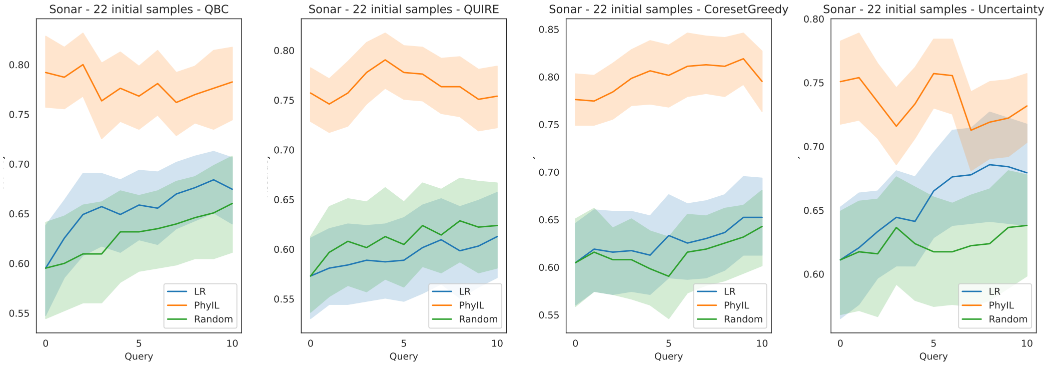
Seeds - 21 initial samples - Uncertainty

Seeds - 21 initial samples - QBC

Seeds - 21 initial samples - CoresetGreedy

(a) (b) (c) (d)

5.2 sonar


Sonar - 22 initial samples - QBC

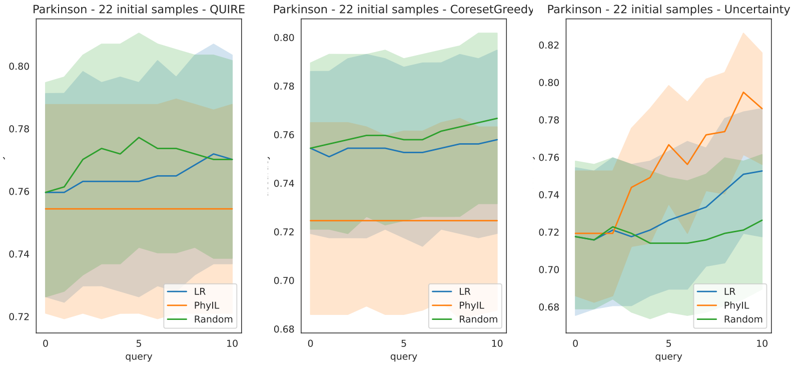
Sonar - 22 initial samples - QUIRE

Sonar - 22 initial samples - CoresetGreedy

Sonar - 22 initial samples - Uncertainty

(a) (b) (c) (d)

5.3 parkinson

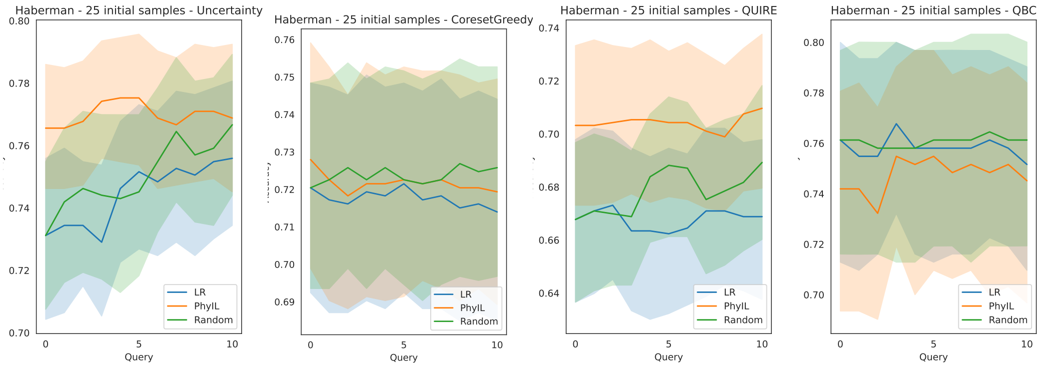

Parkinson - 22 initial samples - QUIRE

Parkinson - 22 initial samples - CoresetGreedy

Parkinson - 22 initial samples - Uncertainty

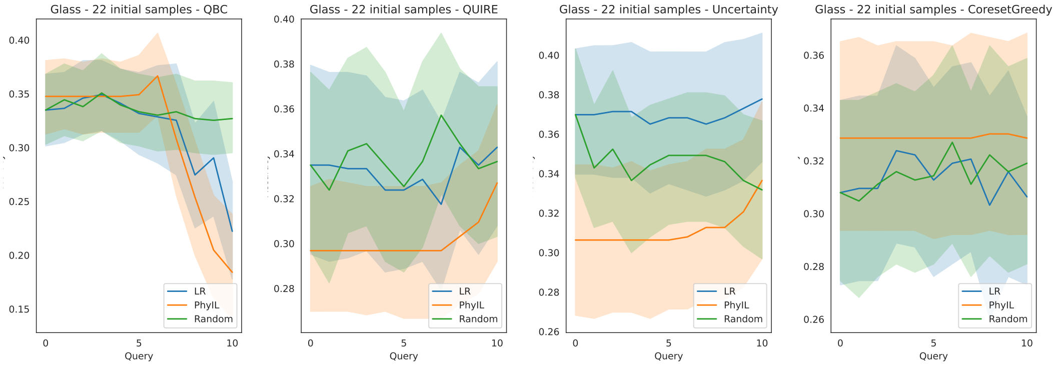
(a) (b) (c)

5.4 haberman



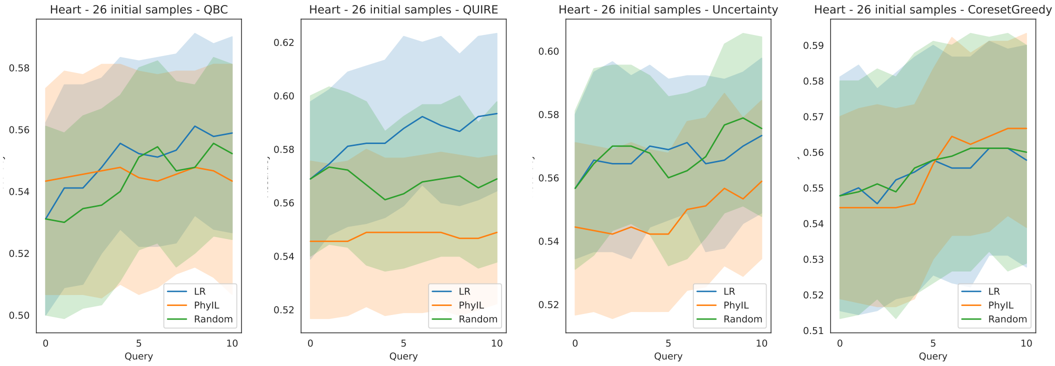
(a) (b) (c) (d)

5.5 glass



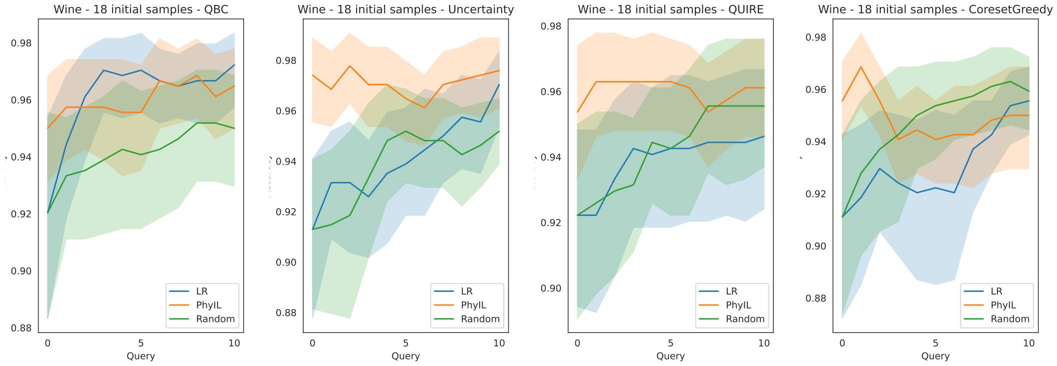
(a) (b) (c) (d)

5.6 heart



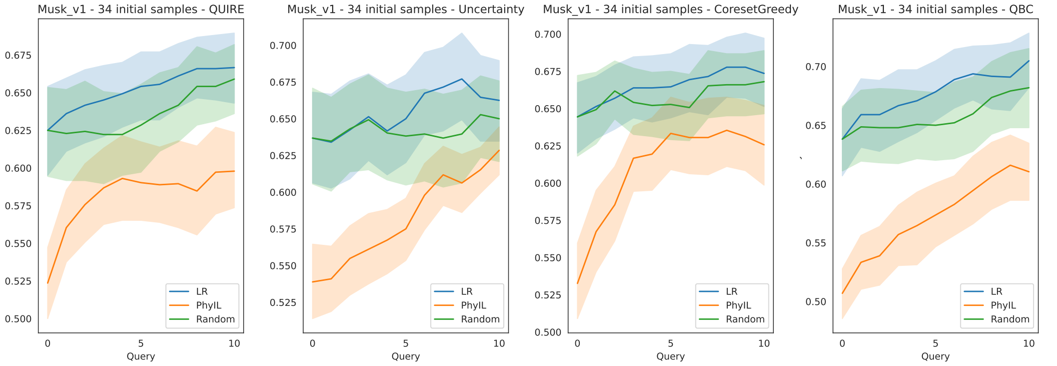
(a) (b) (c) (d)

5.7 wine



(a) (b) (c) (d)

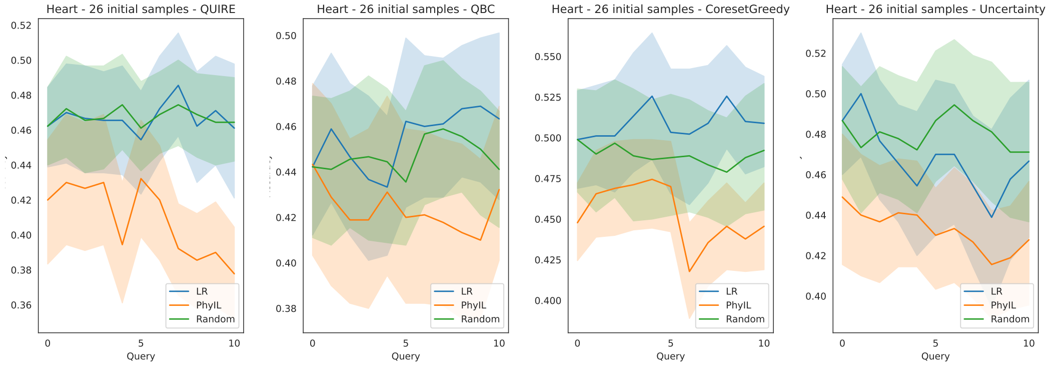
5.8 musk



(a) (b) (c) (d)

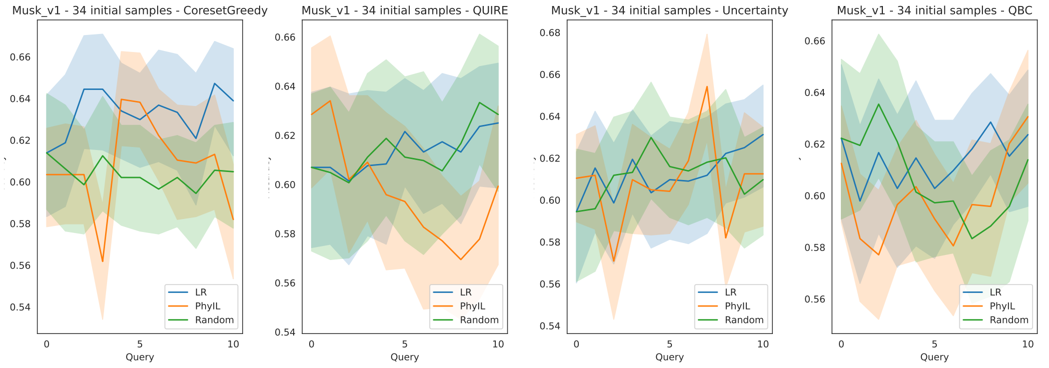
6 Decision Tree

6.1 heart



(a) (b) (c) (d)

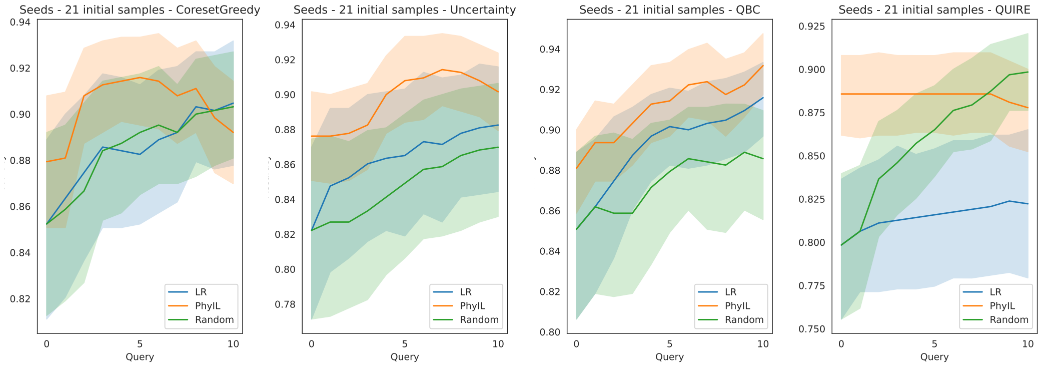
6.2 musk



(a) (b) (c) (d)

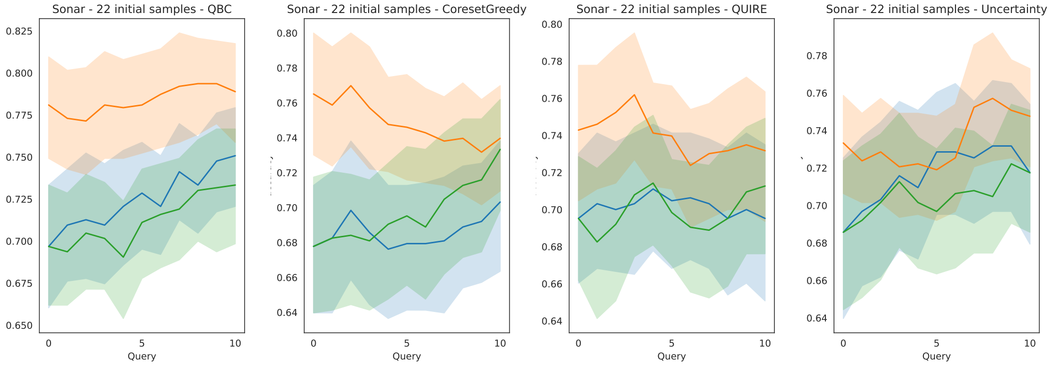
7 MLP

7.1 seeds



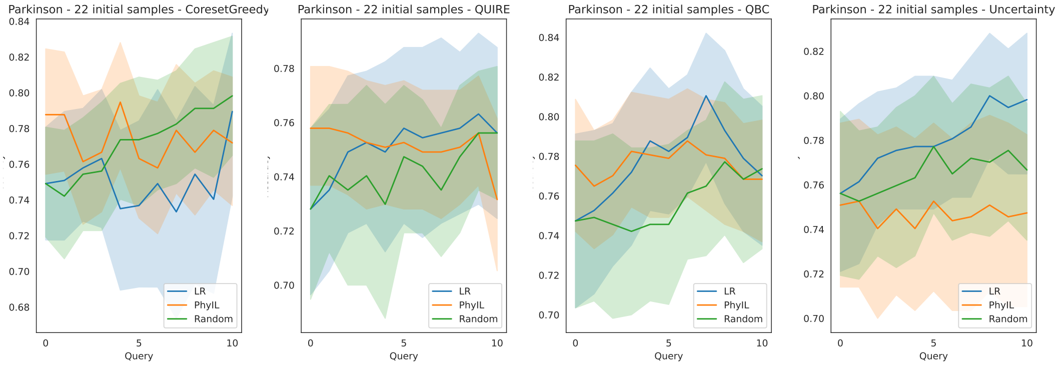
(a) (b) (c) (d)

7.2 sonar



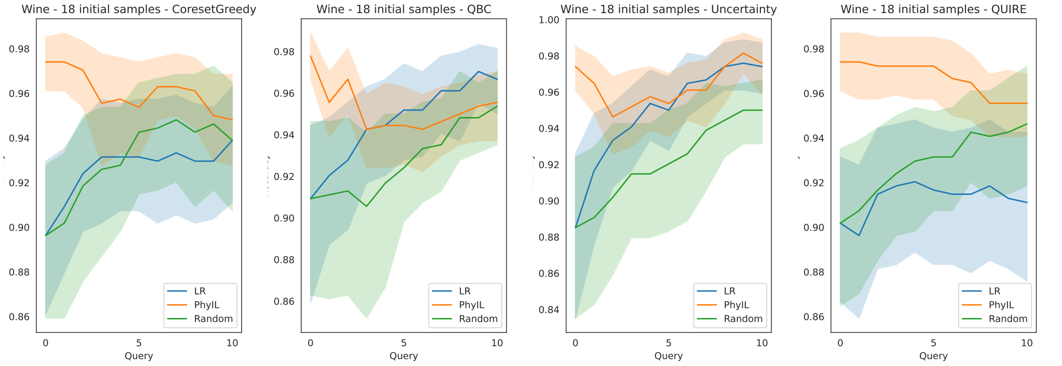
(a) (b) (c) (d)

7.3 parkinson



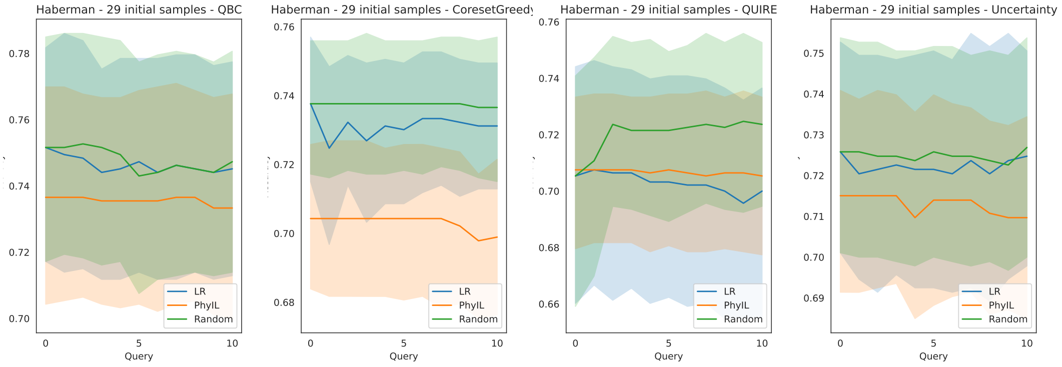
(a) (b) (c) (d)

7.4 wine



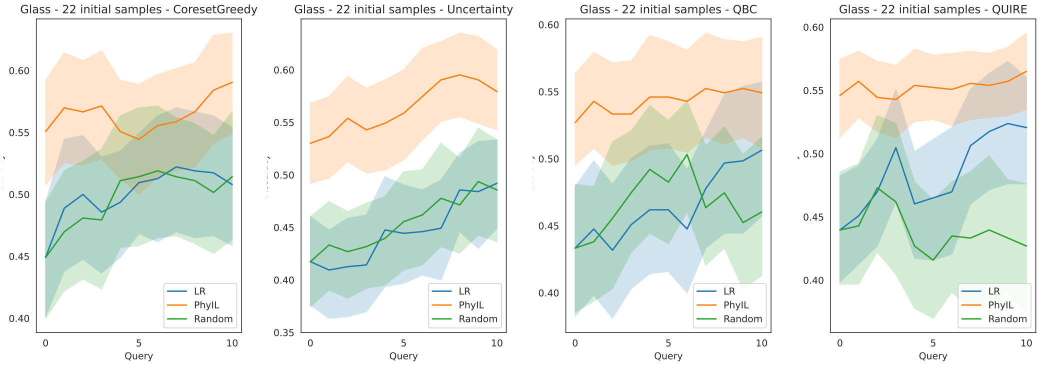
(a) (b) (c) (d)

7.5 haberman



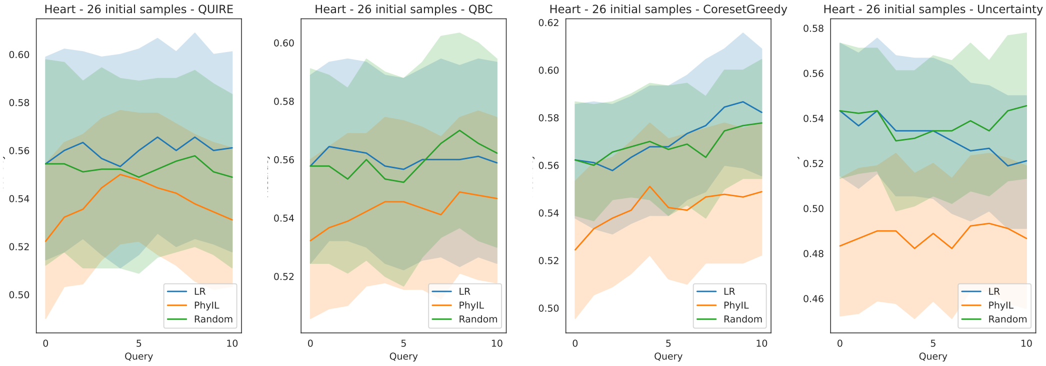
(a) (b) (c) (d)

7.6 glass



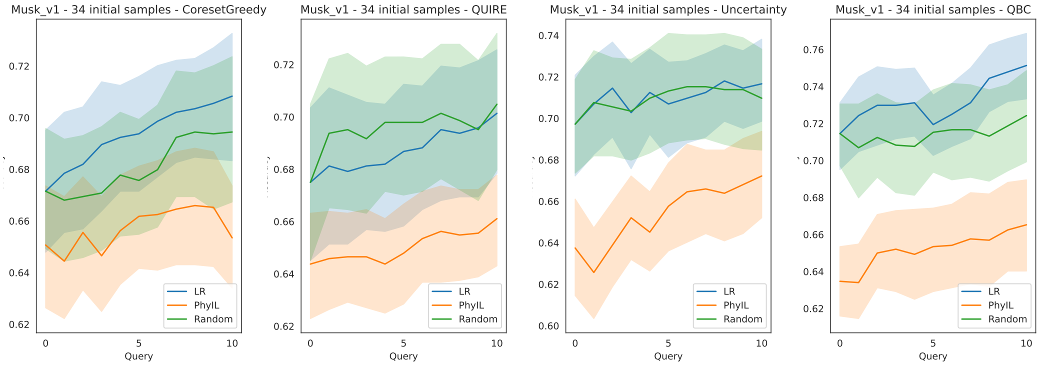
(a) (b) (c) (d)

7.7 heart



(a) (b) (c) (d)

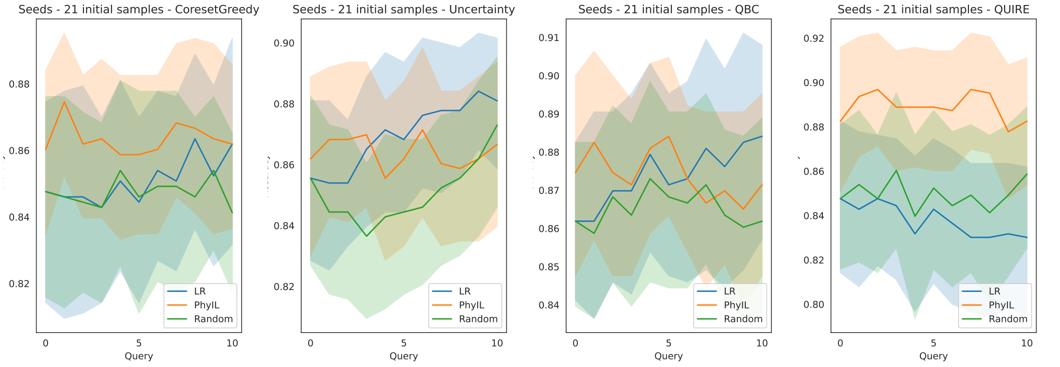
7.8 musk



(a) (b) (c) (d)

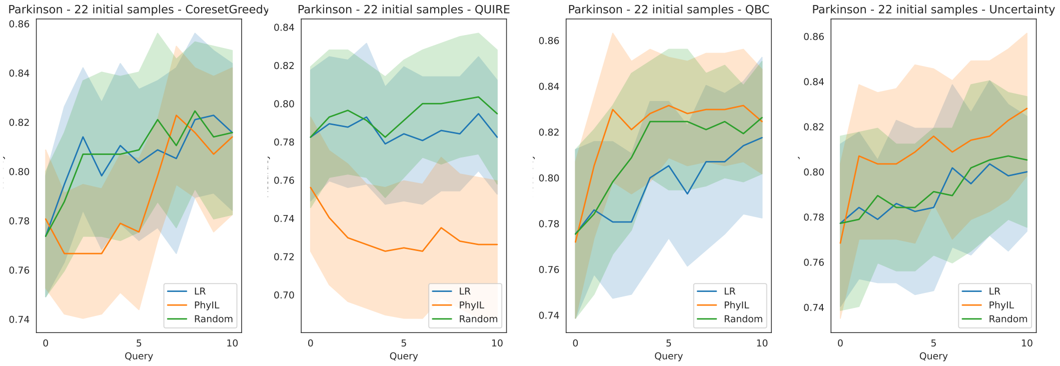
8 Random Forest

8.1 seeds



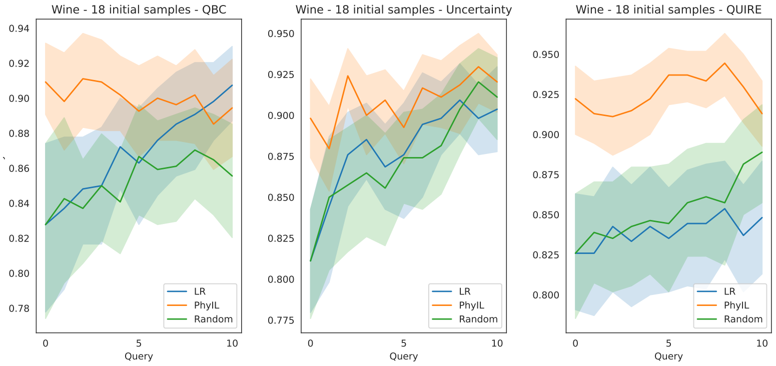
(a) (b) (c) (d)

8.2 parkinson



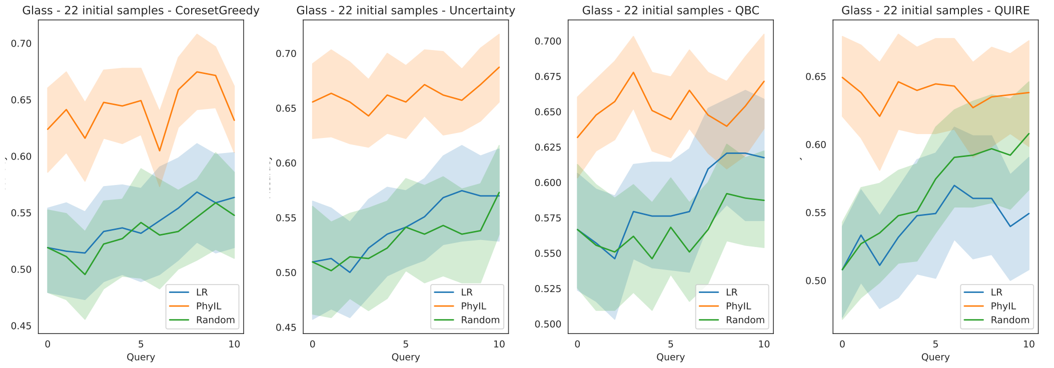
(a) (b) (c) (d)

8.3 wine



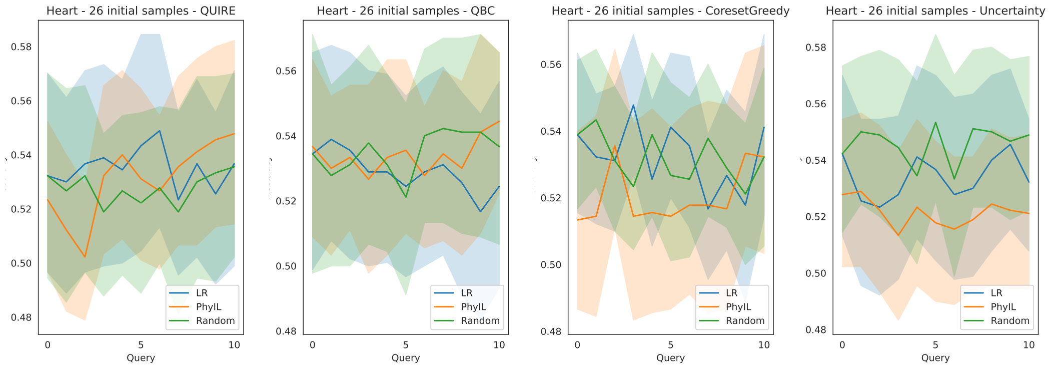
(a) (b) (c)

8.4 glass



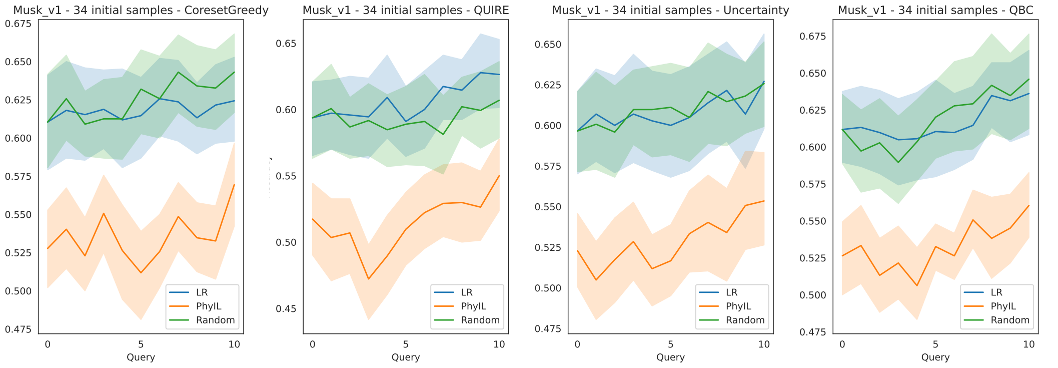
(a) (b) (c) (d)

8.5 heart



(a) (b) (c) (d)

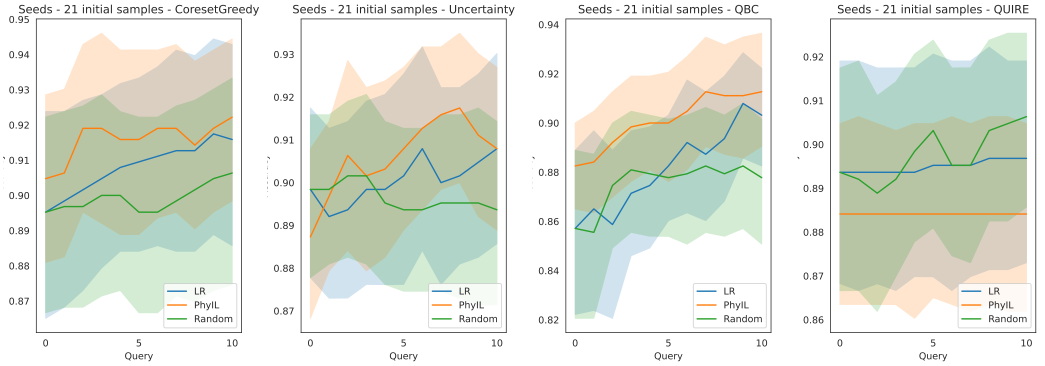
8.6 musk



(a) (b) (c) (d)

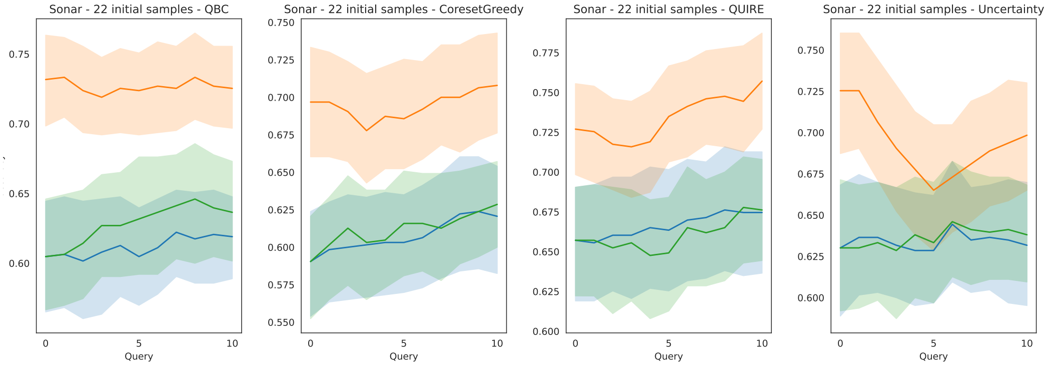
9 KNN

9.1 seeds



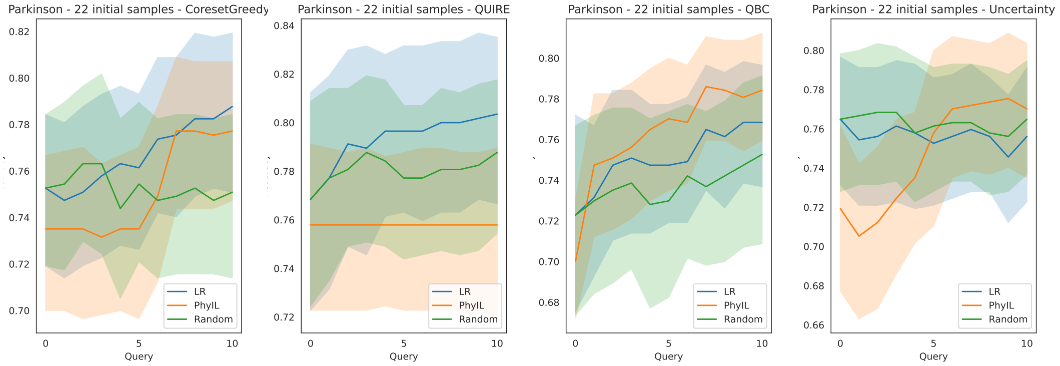
(a) (b) (c) (d)

9.2 sonar



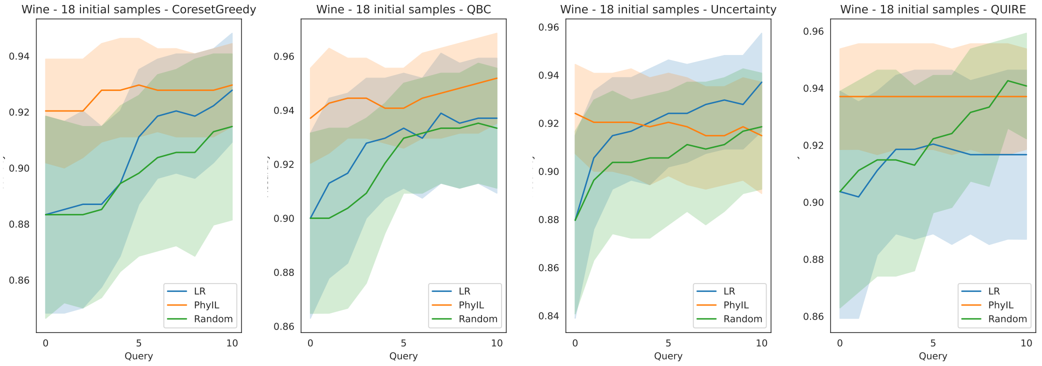
(a) (b) (c) (d)

9.3 parkinson



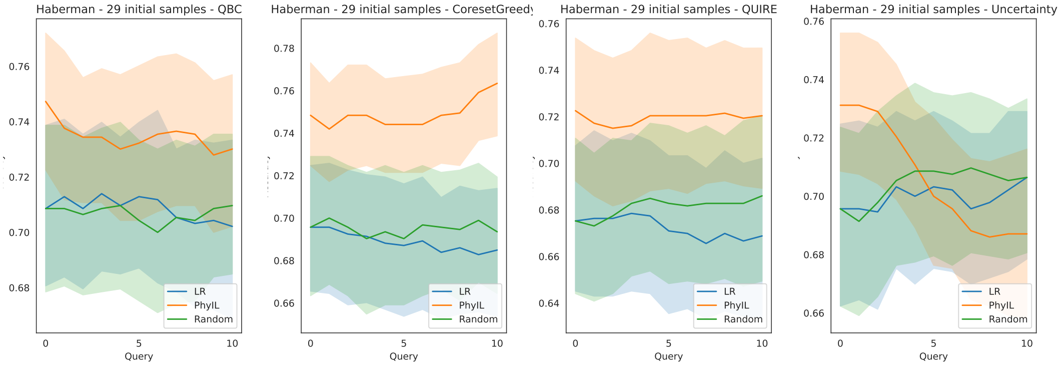
(a) (b) (c) (d)

9.4 wine



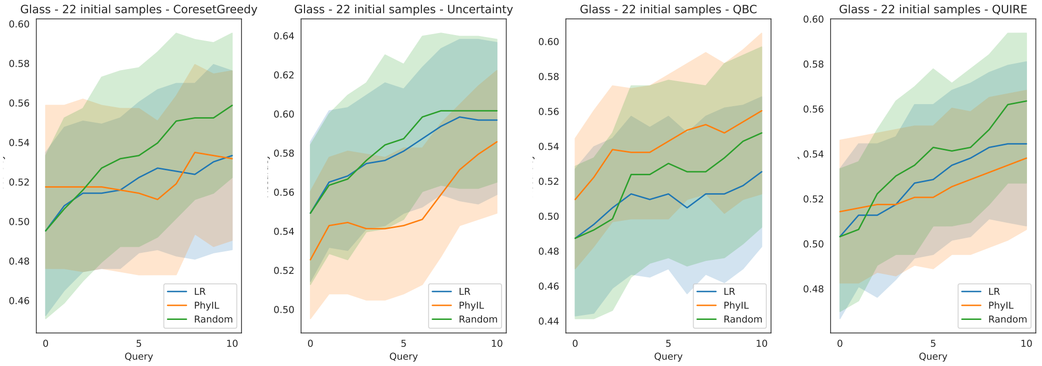
(a) (b) (c) (d)

9.5 haberman



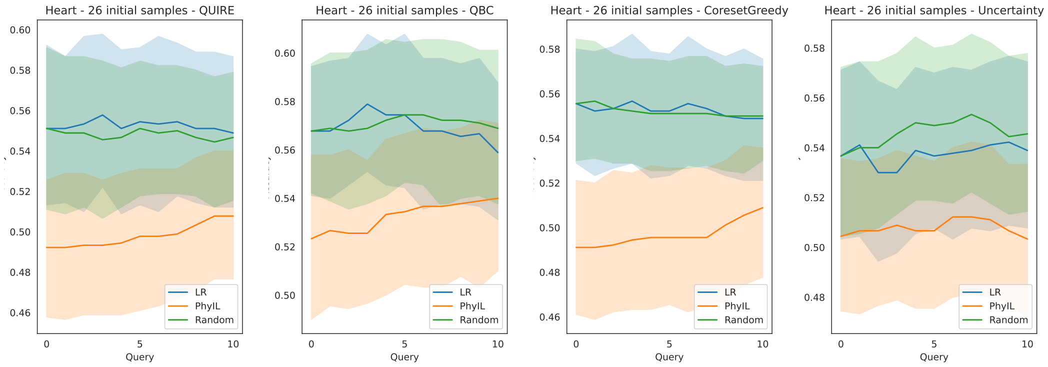
(a) (b) (c) (d)

9.6 glass



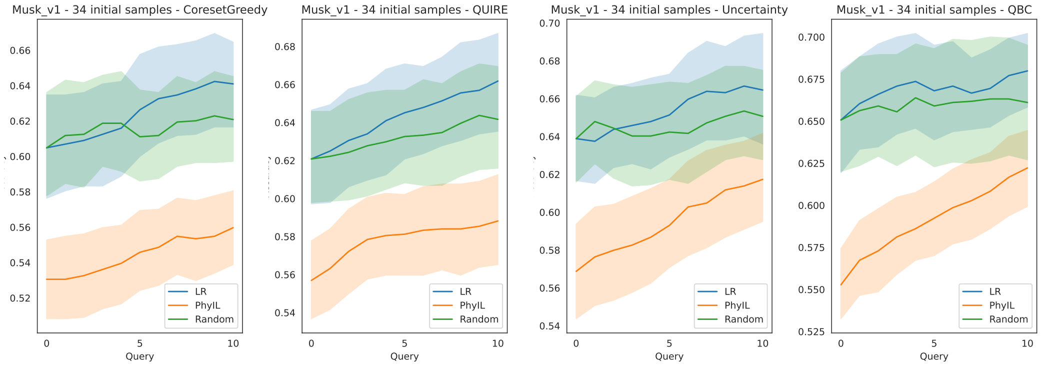
(a) (b) (c) (d)

9.7 heart



(a) (b) (c) (d)

9.8 musk



(a) (b) (c) (d)