

# Supplementary Material

Breno Caetano da Silva  
Jorge Luiz Franco  
Eric K. Tokuda  
Alexandre Delbem

June 29, 2024

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>ADABOOST</b>	<b>3</b>
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
<b>3</b>	<b>Linear Regression</b>	<b>5</b>
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
<b>4</b>	<b>Gaussian Process</b>	<b>8</b>
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
<b>5</b>	<b>SVM</b>	<b>10</b>
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

<b>6</b>	<b>Decision Tree</b>	<b>13</b>
6.1	heart . . . . .	13
6.2	musk . . . . .	13
<b>7</b>	<b>MLP</b>	<b>14</b>
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
<b>8</b>	<b>Random Forest</b>	<b>16</b>
8.1	seeds . . . . .	16
8.2	parkinson . . . . .	17
8.3	wine . . . . .	17
8.4	glass . . . . .	17
8.5	heart . . . . .	18
8.6	musk . . . . .	18
<b>9</b>	<b>KNN</b>	<b>18</b>
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.

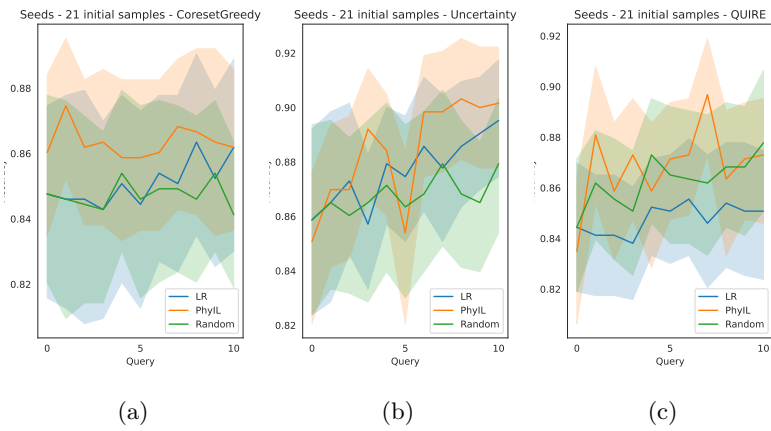
We evaluated PhyIL on nine diverse real-world datasets: seeds, sonar, parkinson, wine, haberman, glass, heart, and musk. These datasets were chosen for their variety in features and challenges, offering a comprehensive test of PhyIL’s effectiveness. For example, the seeds dataset involves grain classification, 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.

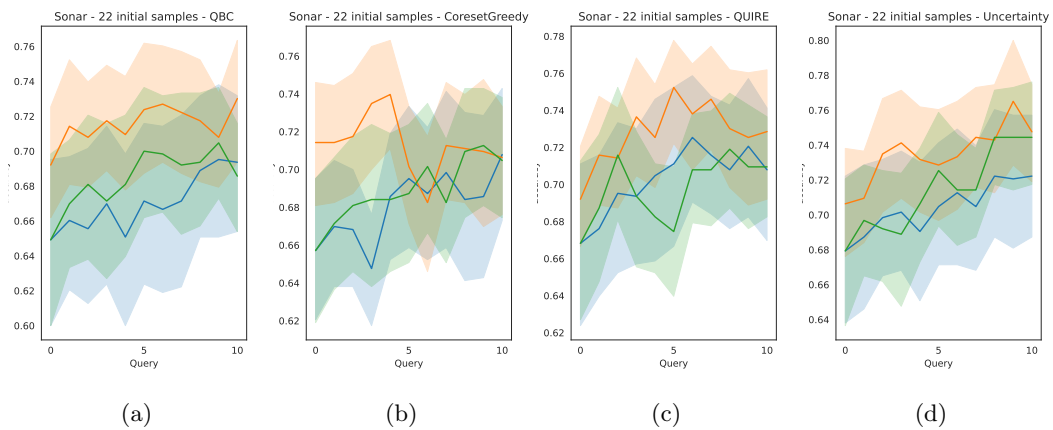
Our results show that PhyIL outperforms these methods, leading to better initial sample selection and improved model performance. This document provides detailed results for various classifiers, including ADABOOST, Linear Regression, Gaussian Process, Support Vector Machine (SVM), Decision Tree, Multi-Layer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

## 2 ADABOOST

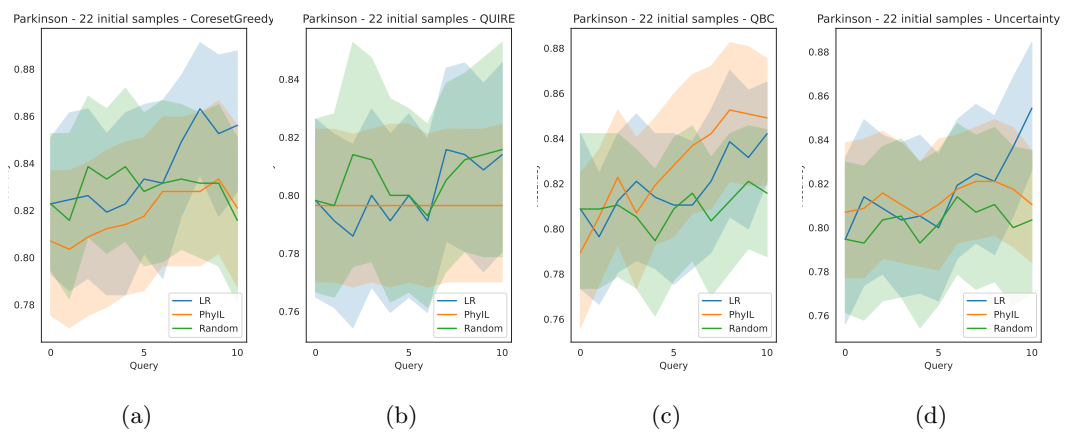
### 2.1 seeds



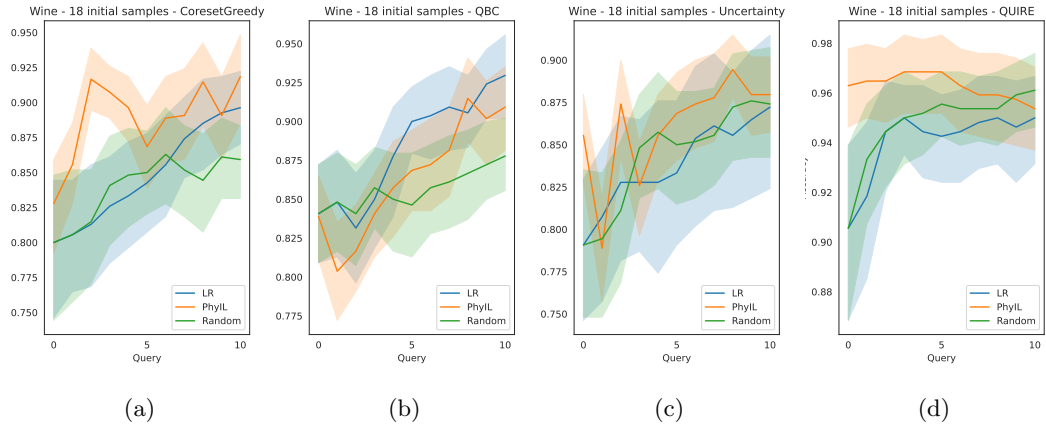
### 2.2 sonar



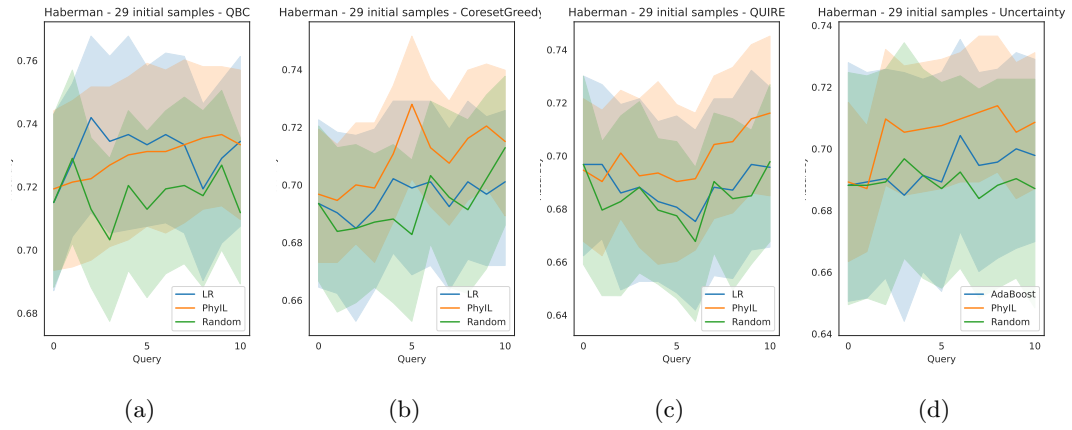
### 2.3 parkinson



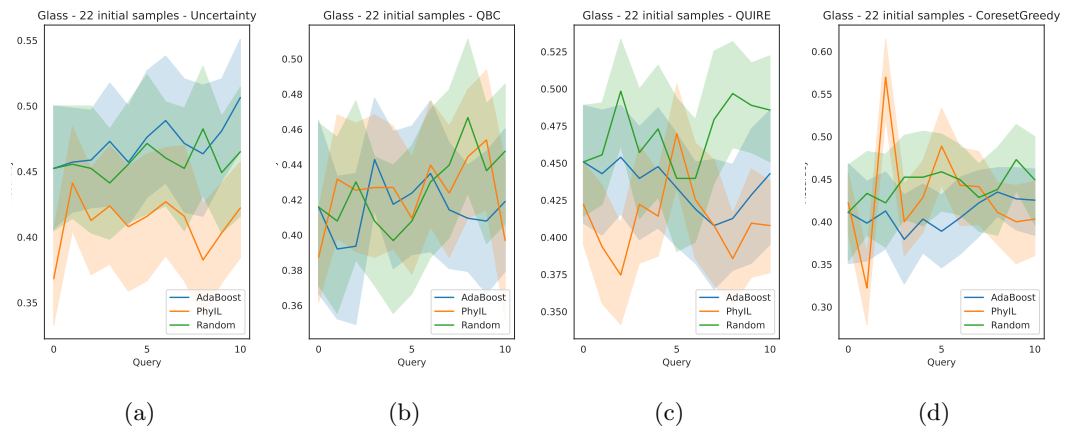
## 2.4 wine



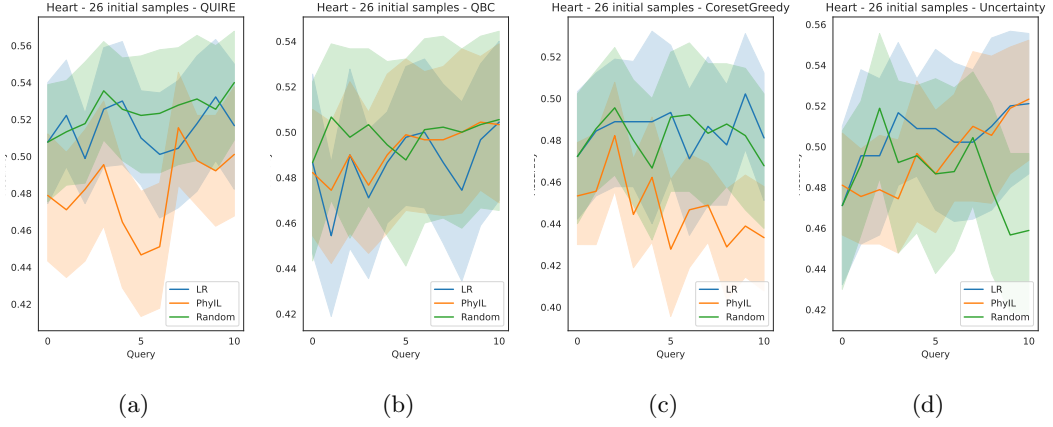
## 2.5 haberman



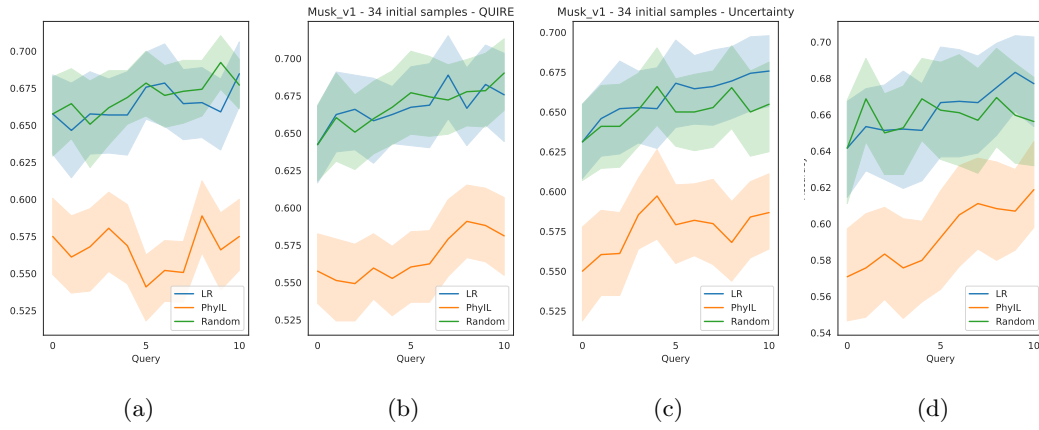
## 2.6 glass



## 2.7 heart

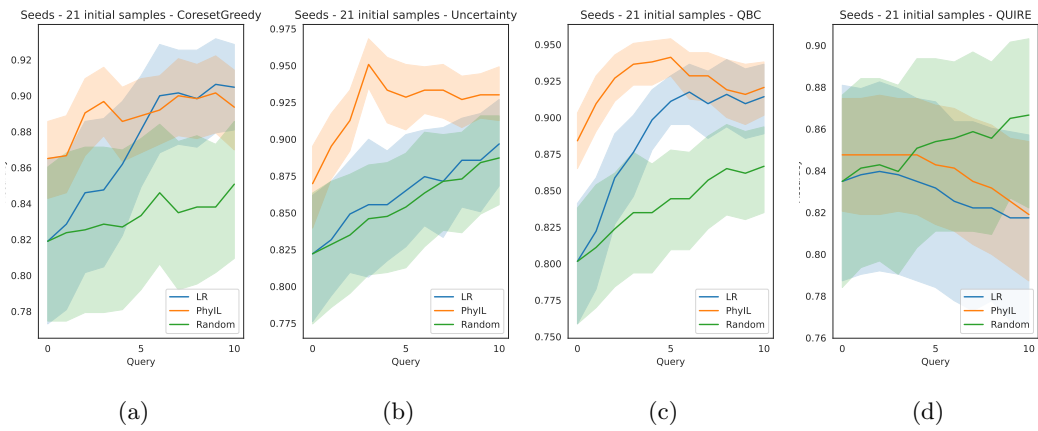


## 2.8 musk

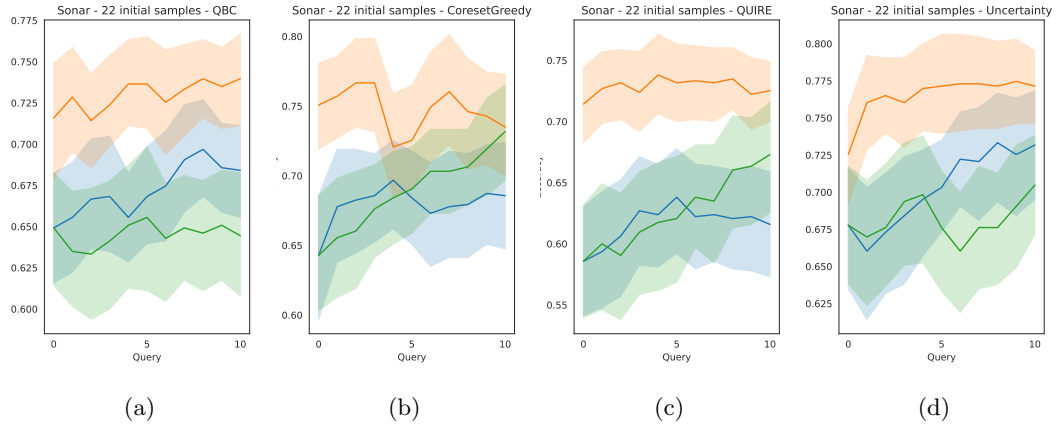


## 3 Linear Regression

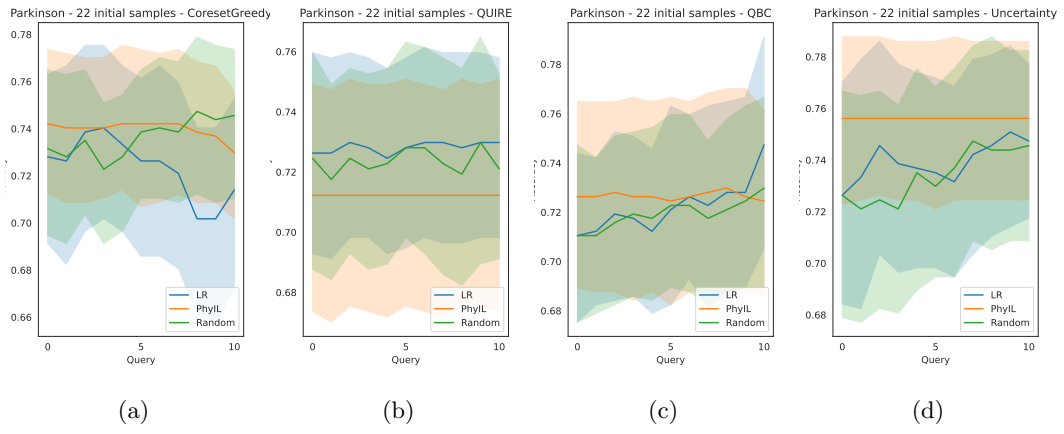
### 3.1 seeds



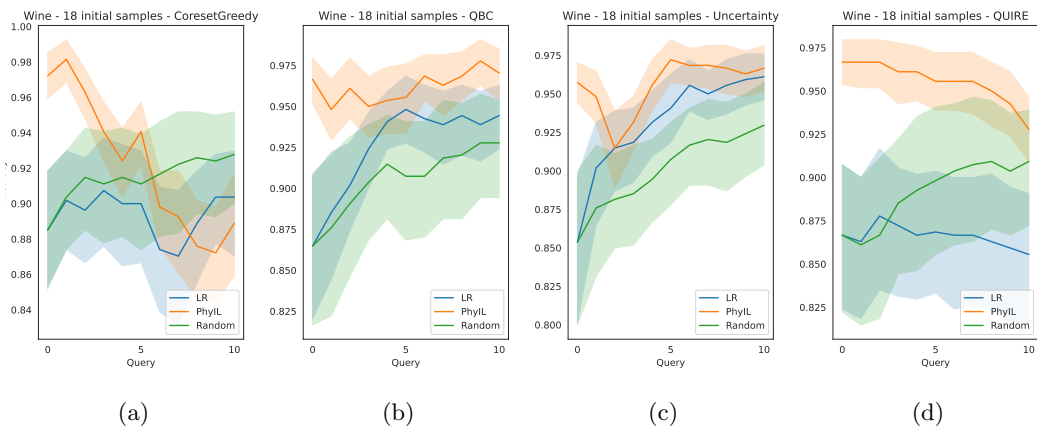
### 3.2 sonar



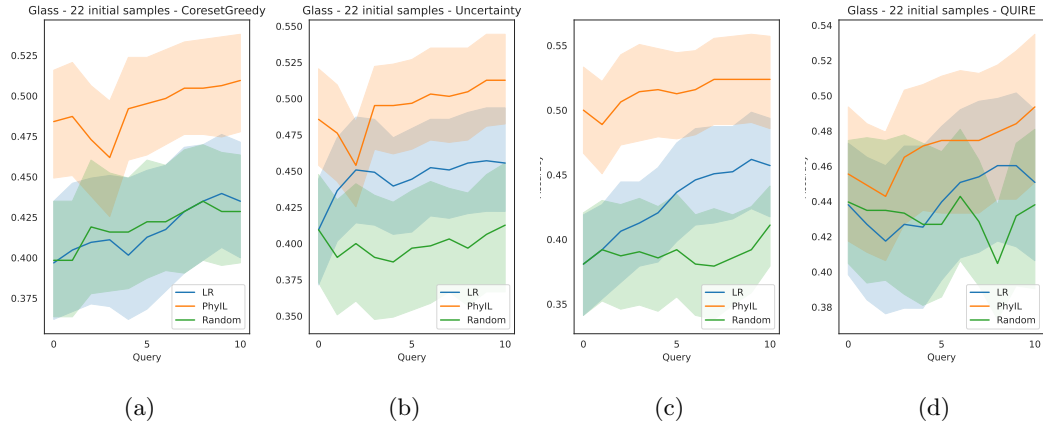
### 3.3 parkinson



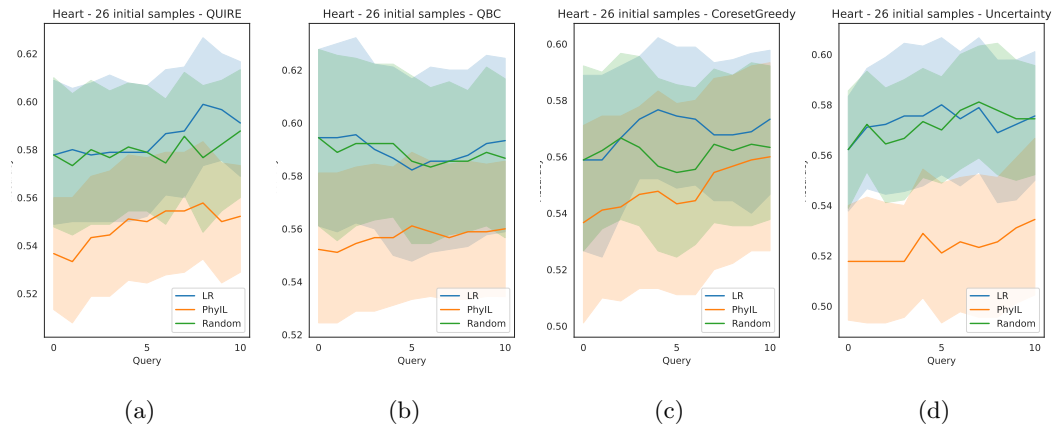
### 3.4 wine



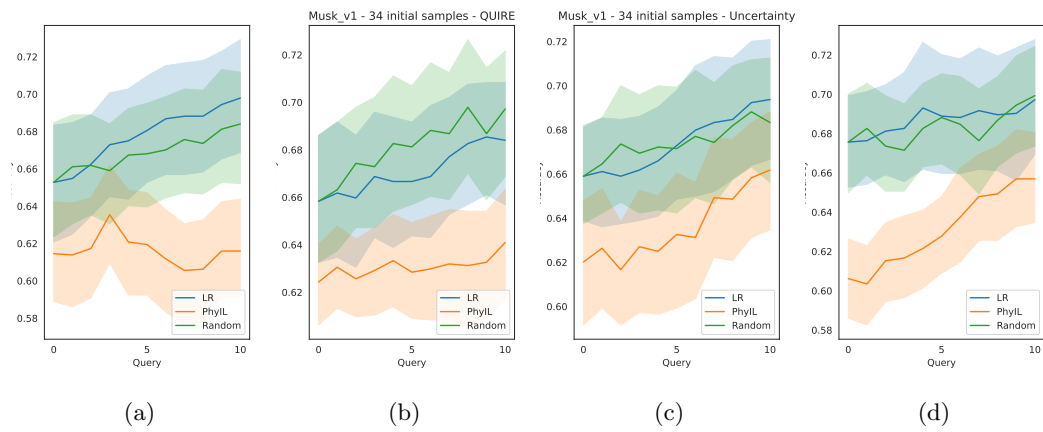
### 3.5 glass



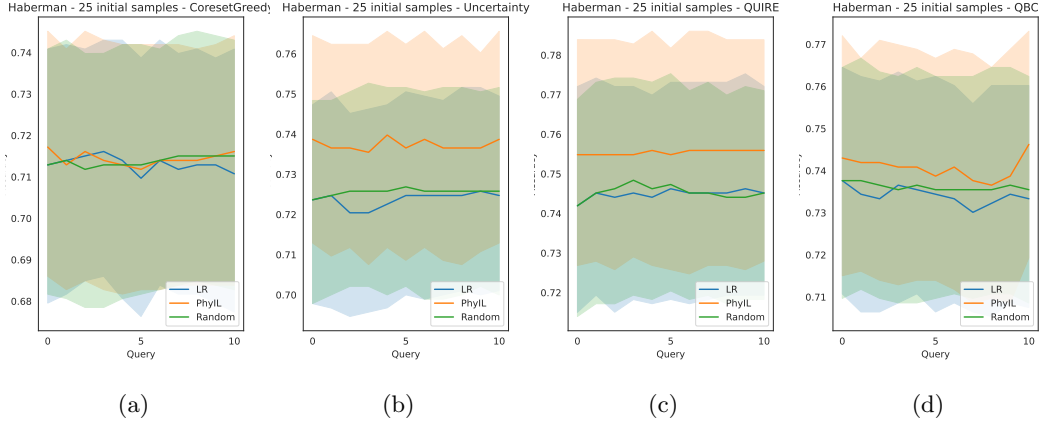
### 3.6 heart



### 3.7 musk

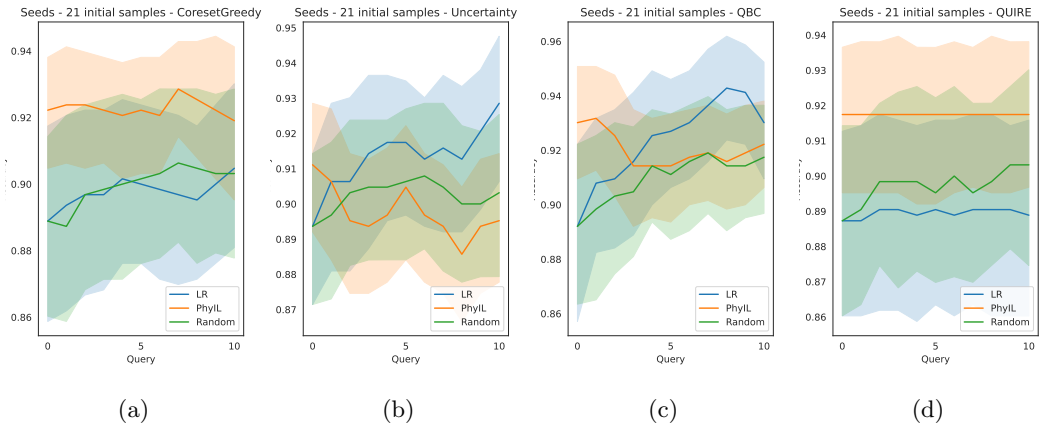


### 3.8 haberman

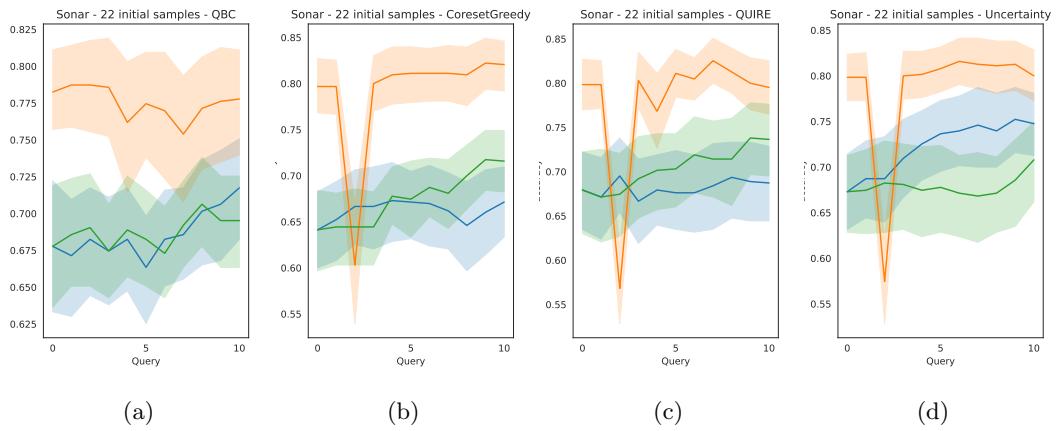


## 4 Gaussian Process

### 4.1 seeds

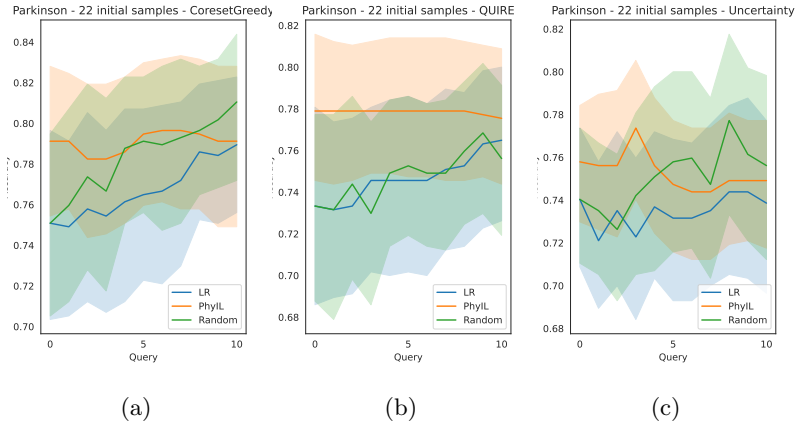


### 4.2 sonar

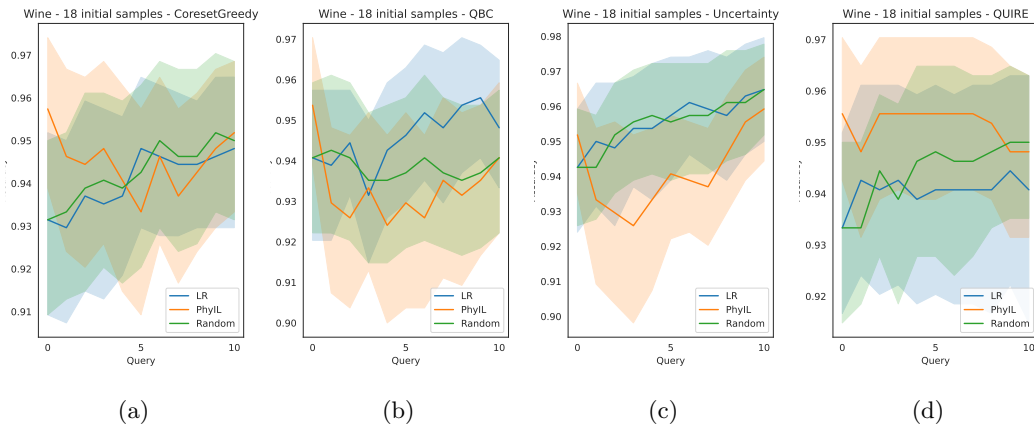


### 4.3 parkinson

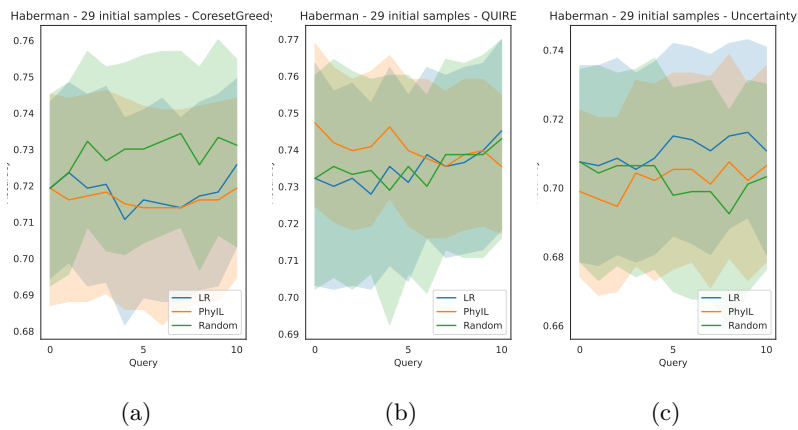




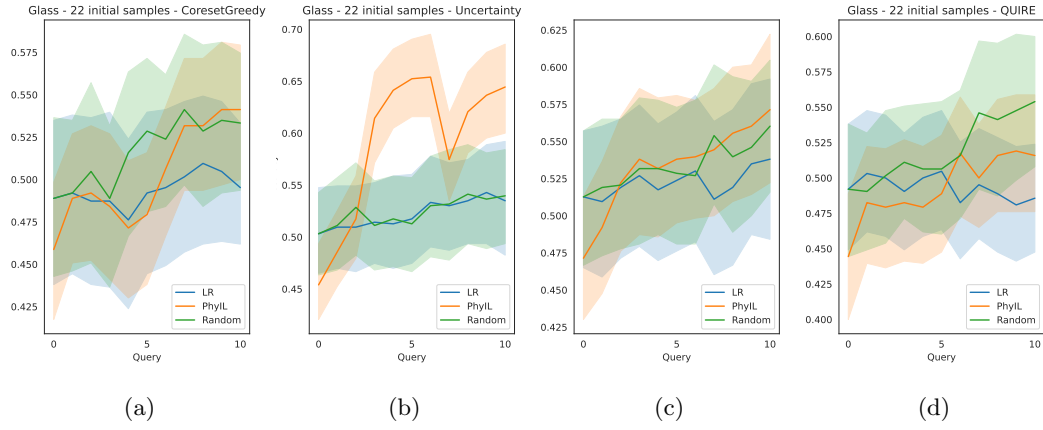
#### 4.4 wine



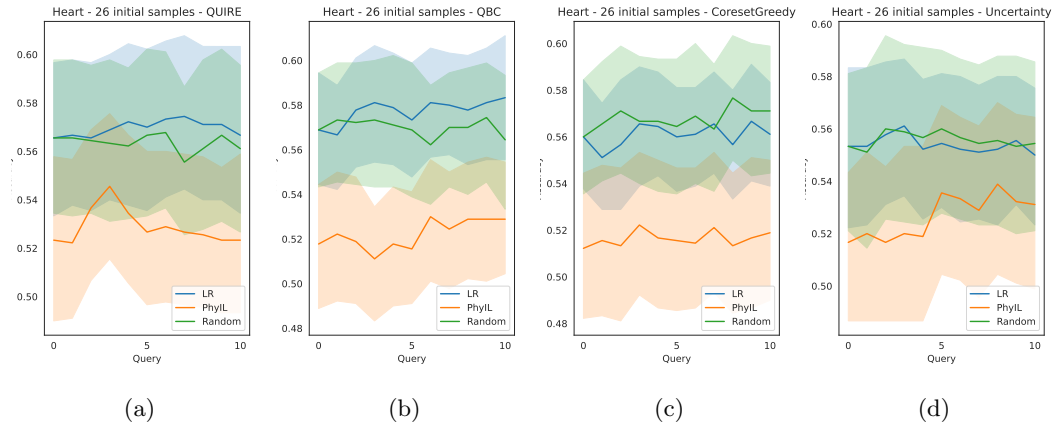
#### 4.5 haberman



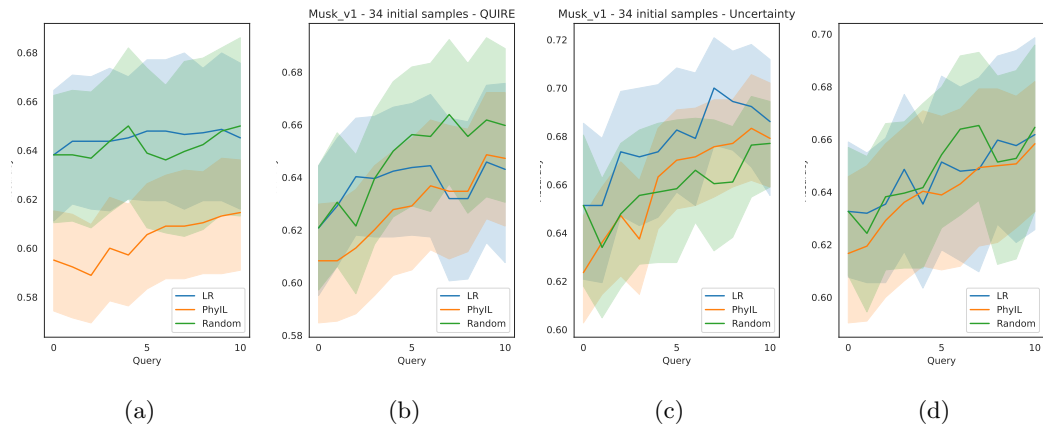
#### 4.6 glass



## 4.7 heart

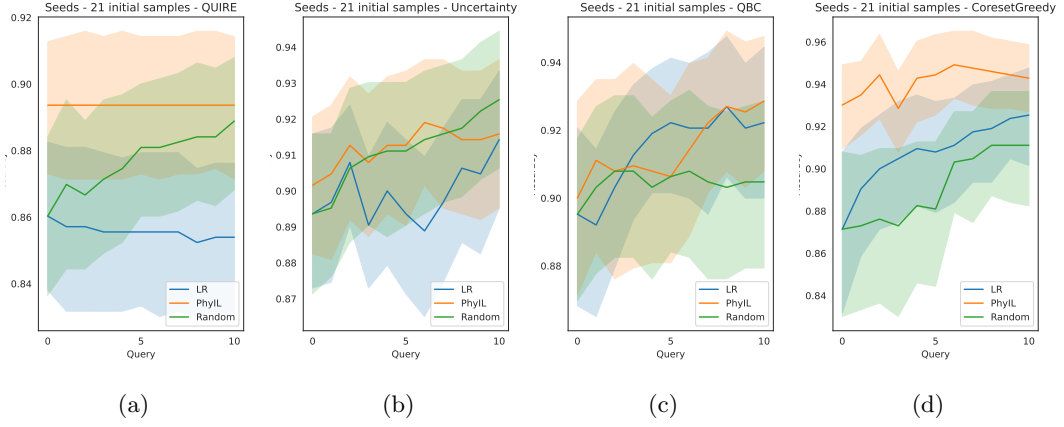


## 4.8 musk

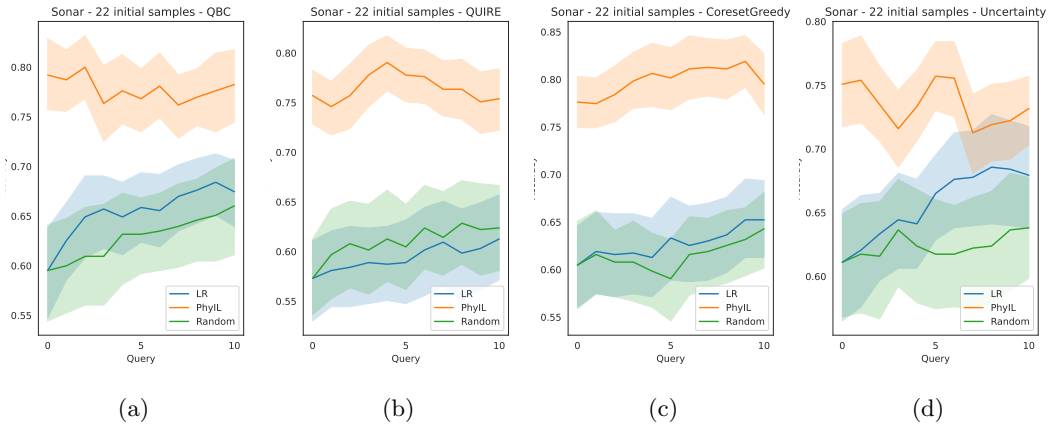


## 5 SVM

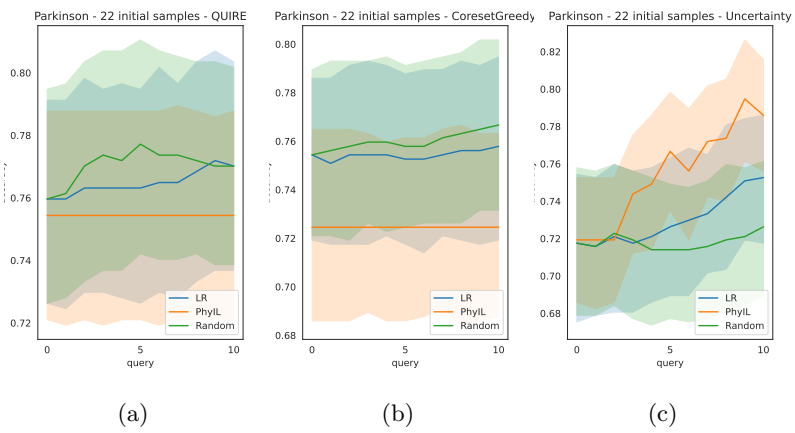
### 5.1 seeds



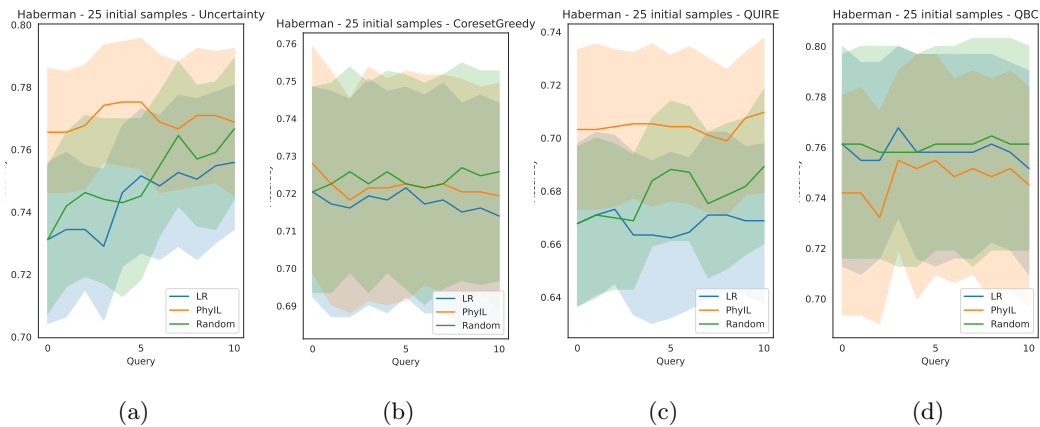
## 5.2 sonar



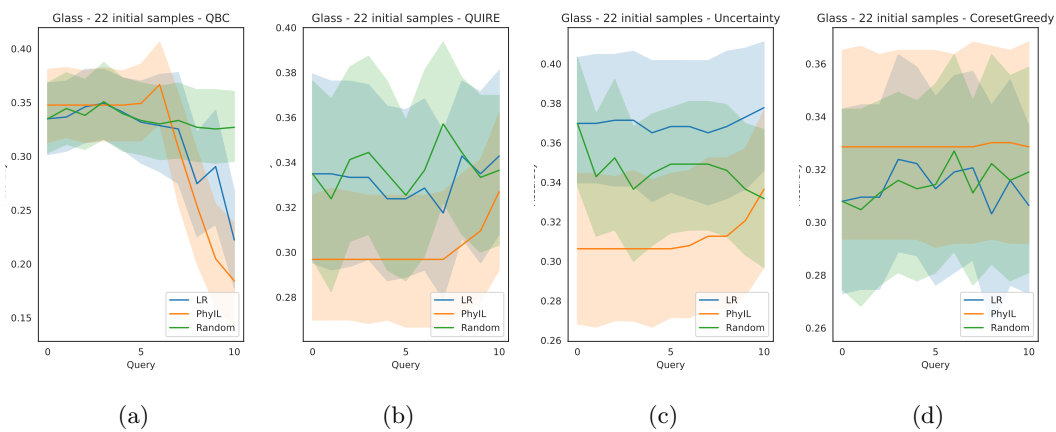
## 5.3 parkinson



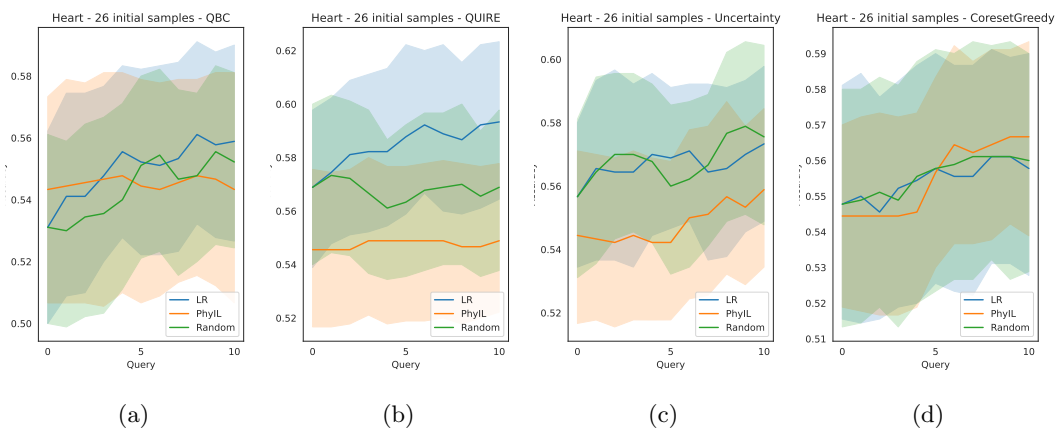
## 5.4 haberman



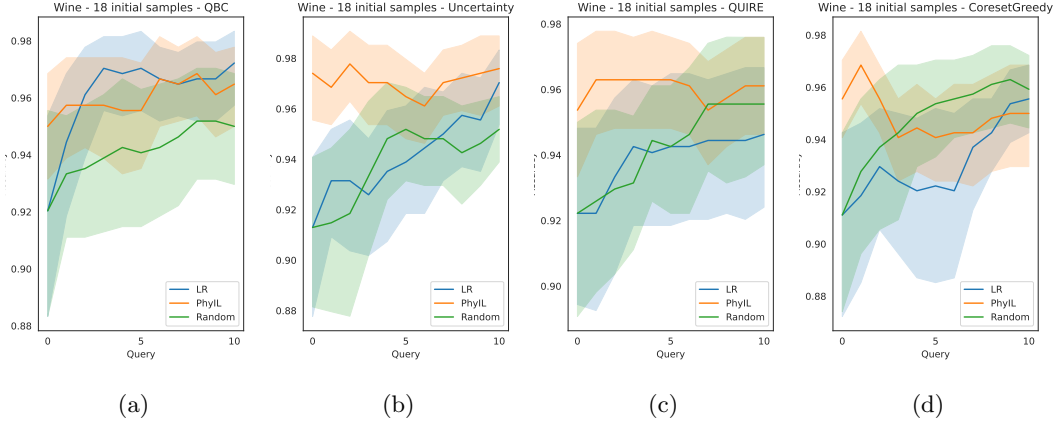
## 5.5 glass



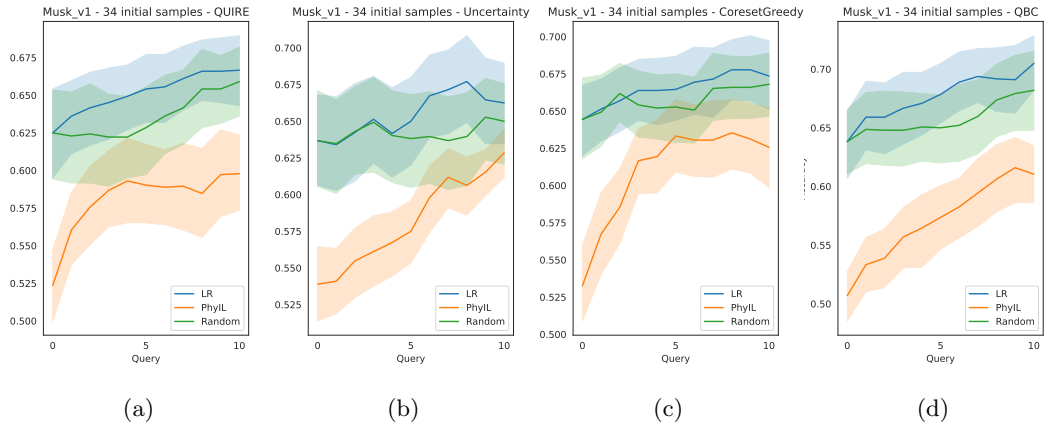
## 5.6 heart



## 5.7 wine

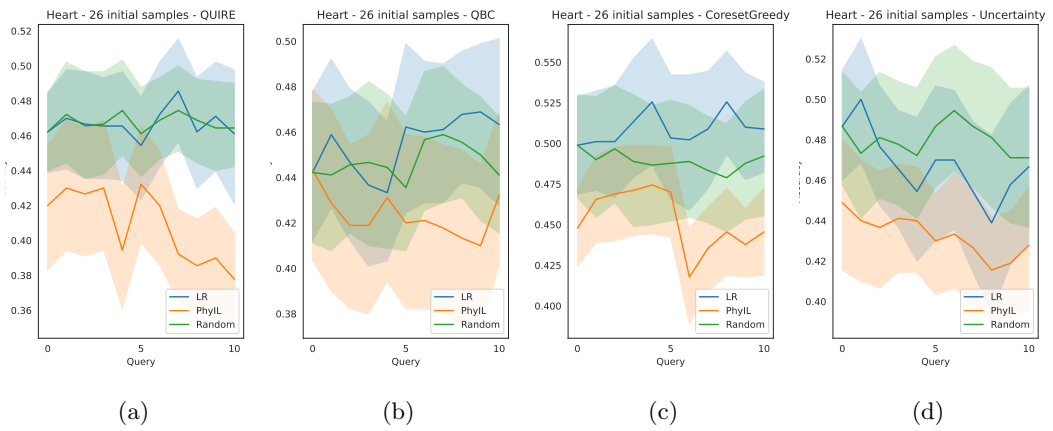


## 5.8 musk

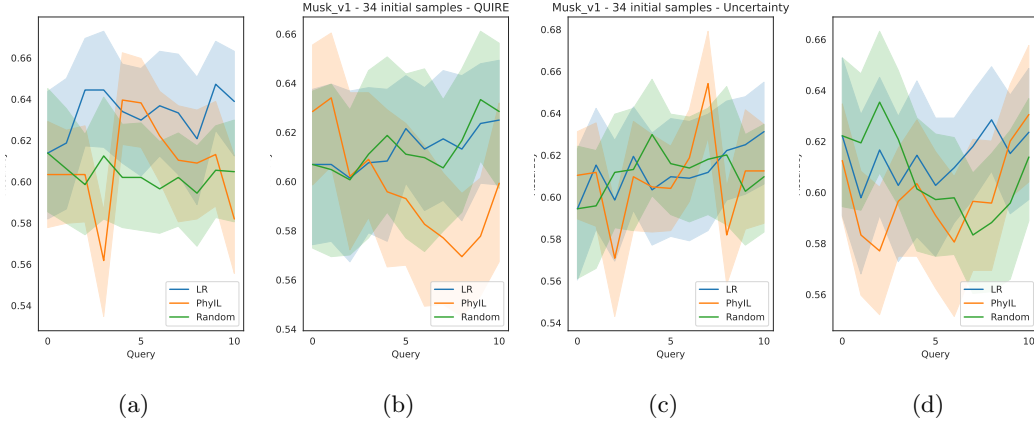


## 6 Decision Tree

### 6.1 heart

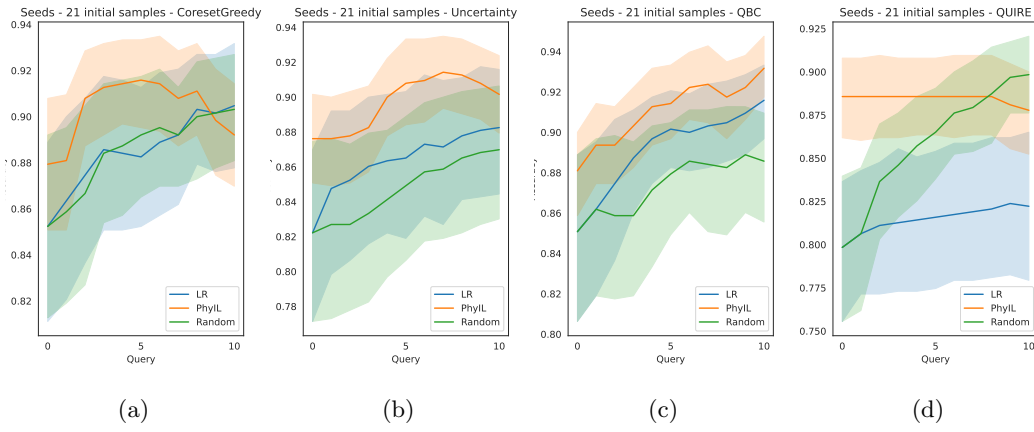


### 6.2 musk

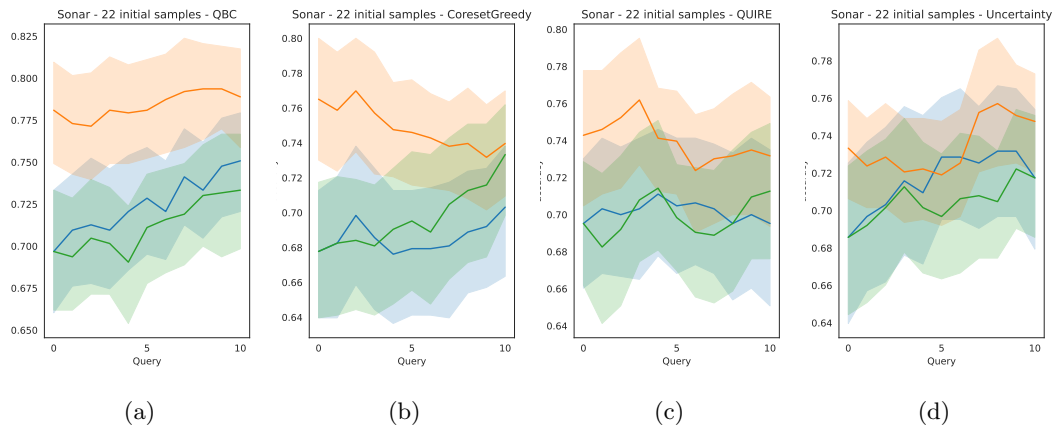


## 7 MLP

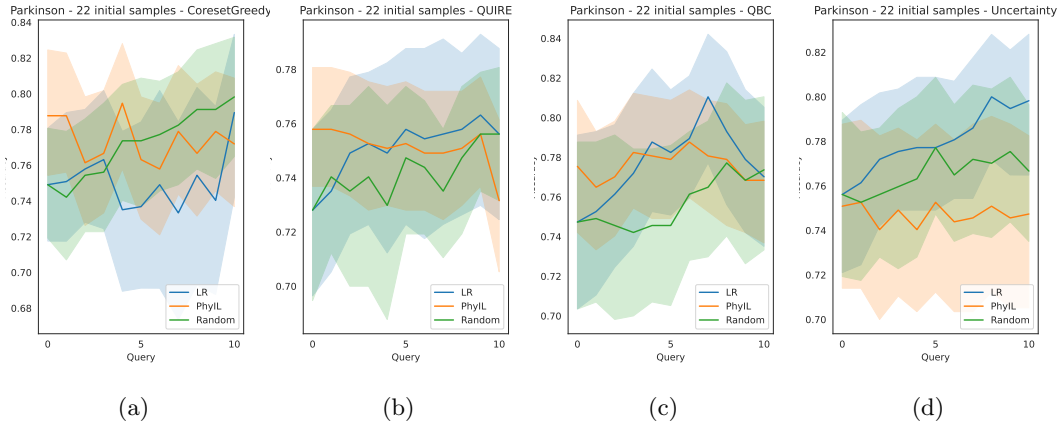
### 7.1 seeds



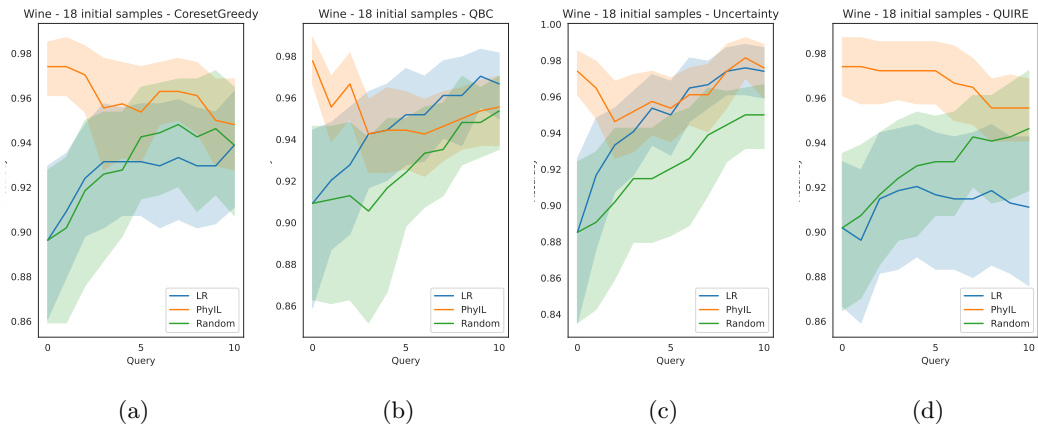
### 7.2 sonar



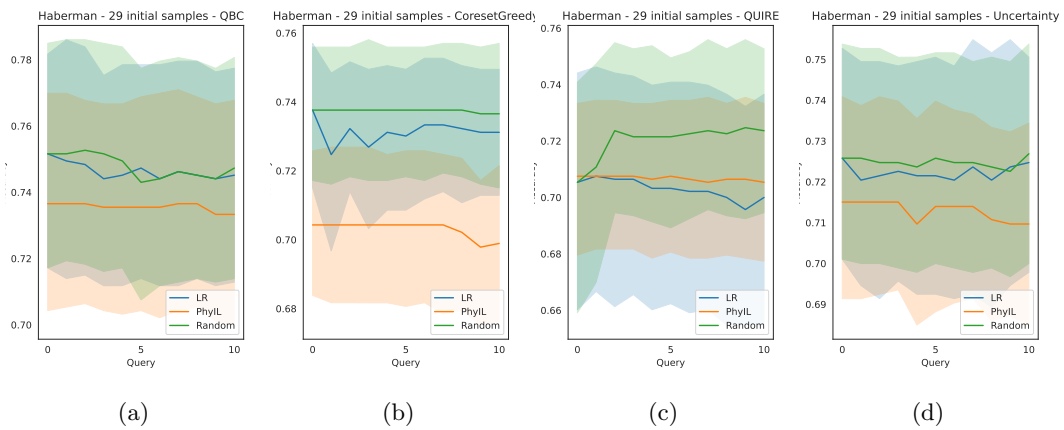
### 7.3 parkinson



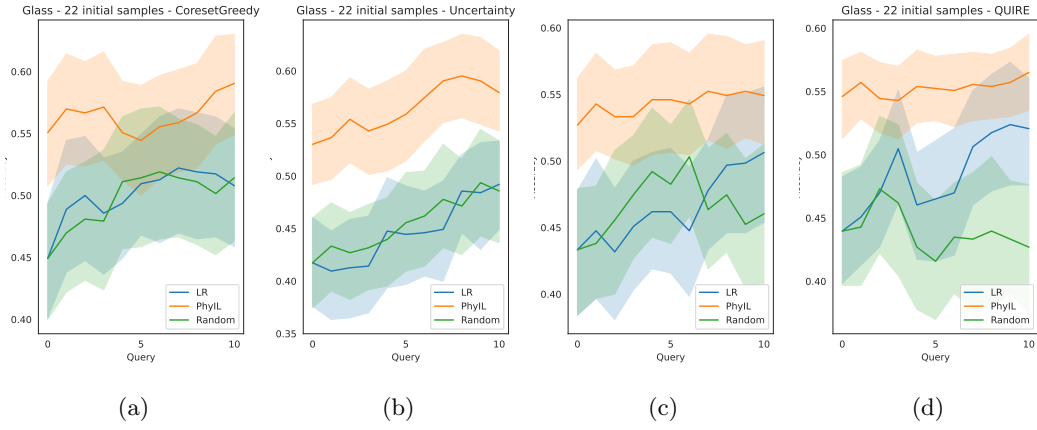
## 7.4 wine



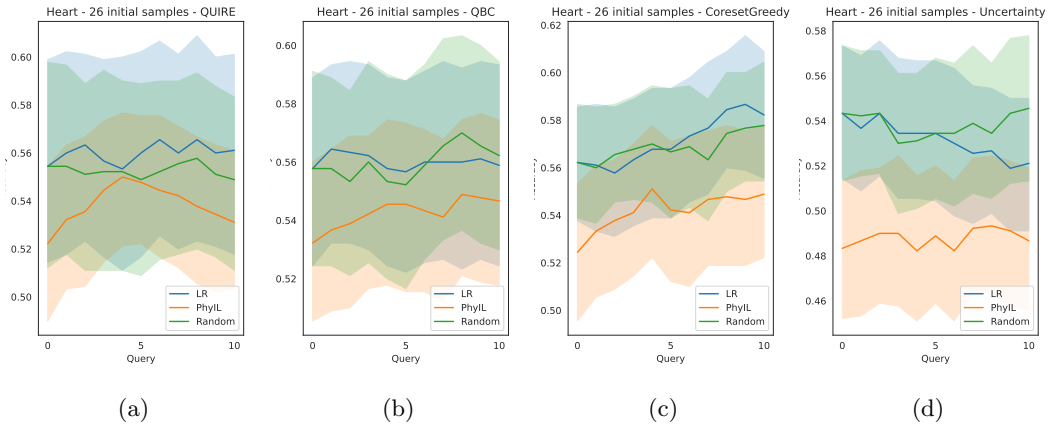
## 7.5 haberman



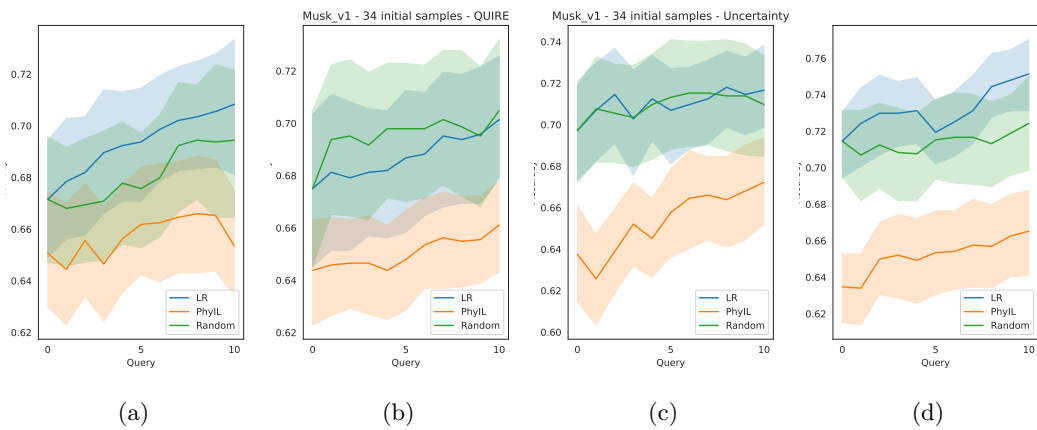
## 7.6 glass



## 7.7 heart



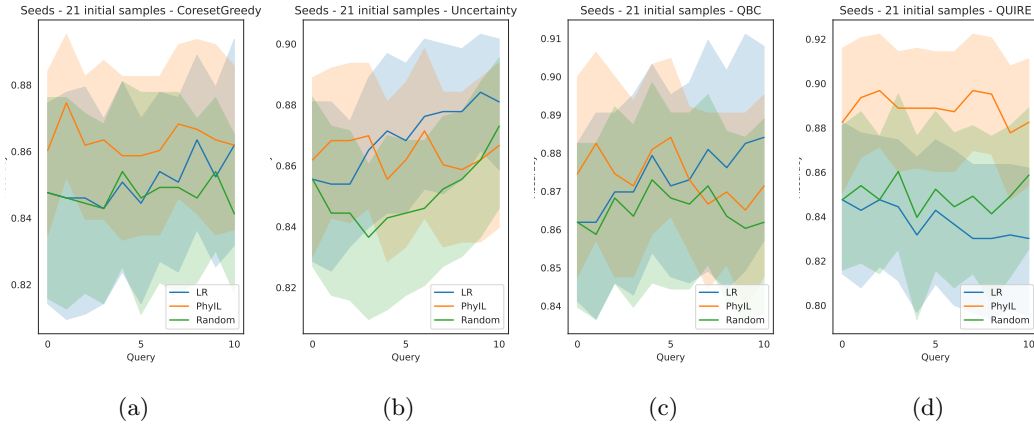
## 7.8 musk



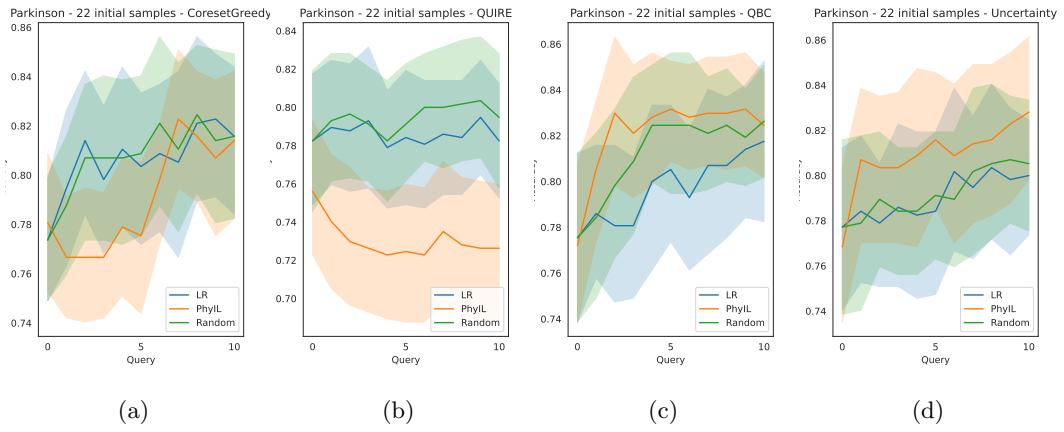
## 8 Random Forest

### 8.1 seeds

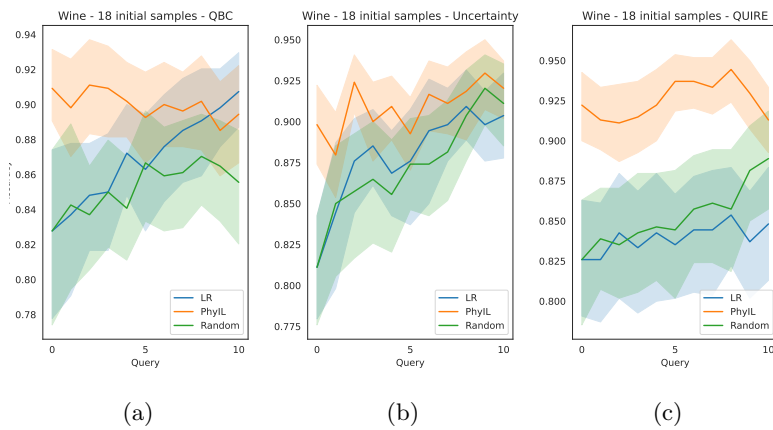




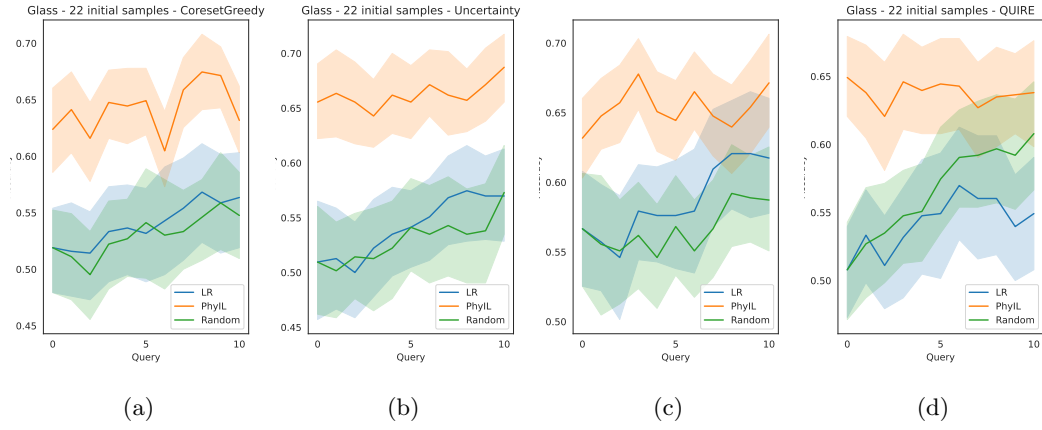
## 8.2 parkinson



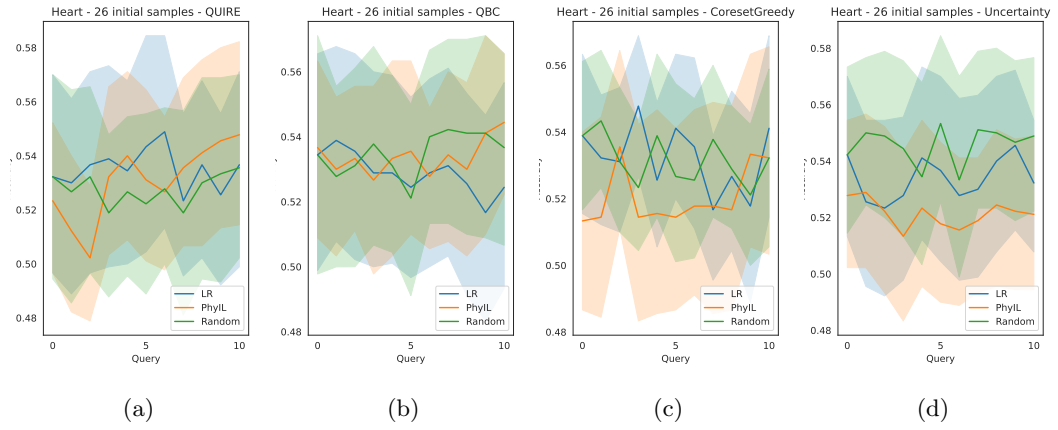
## 8.3 wine



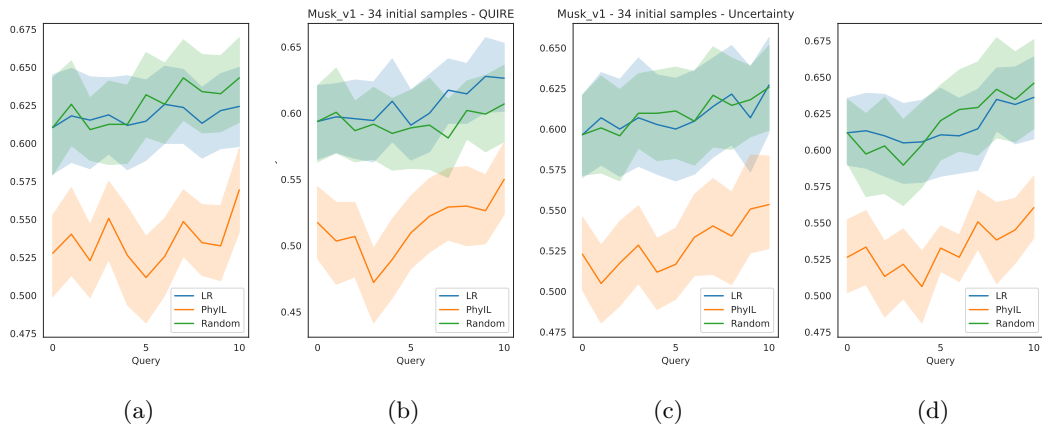
## 8.4 glass



## 8.5 heart

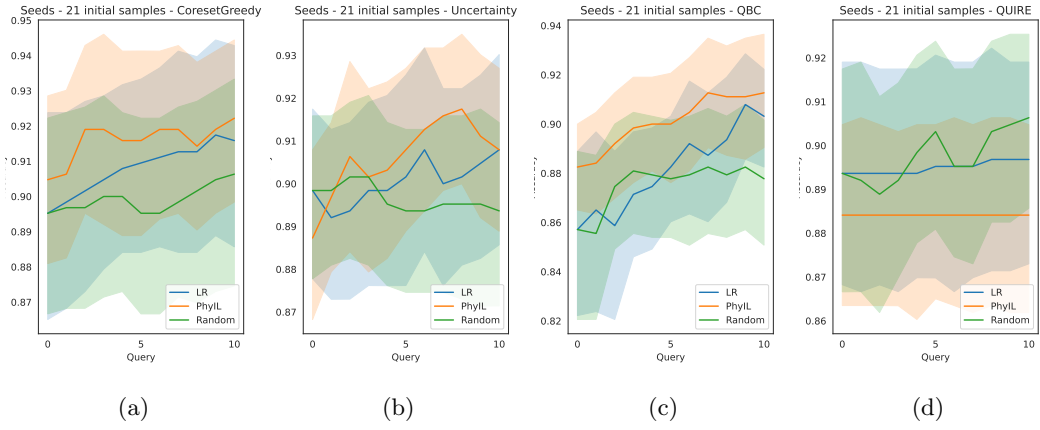


## 8.6 musk

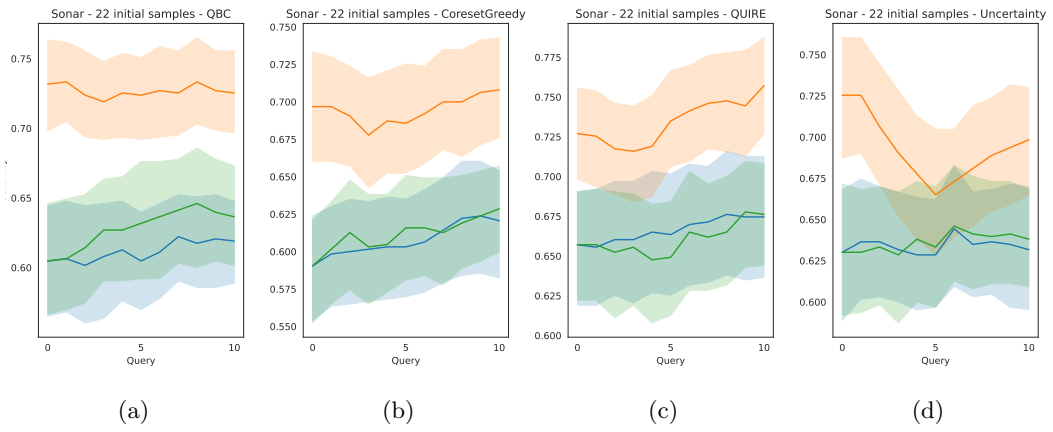


## 9 KNN

### 9.1 seeds



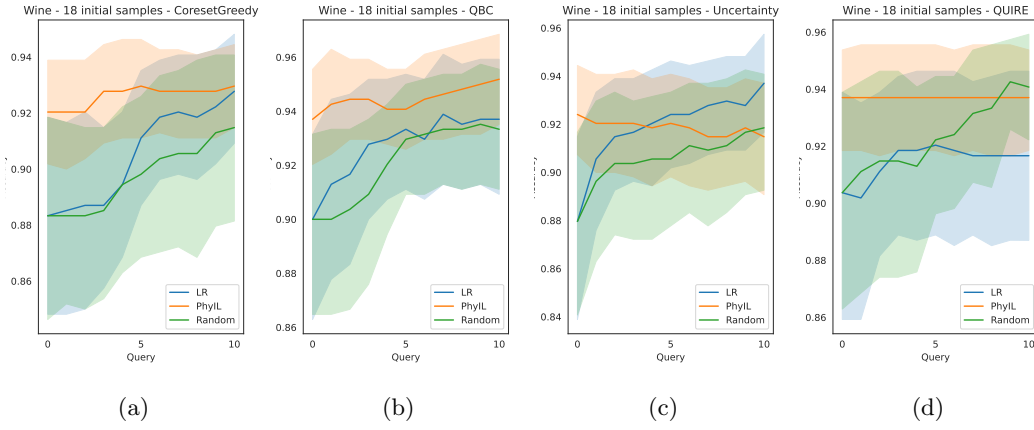
## 9.2 sonar



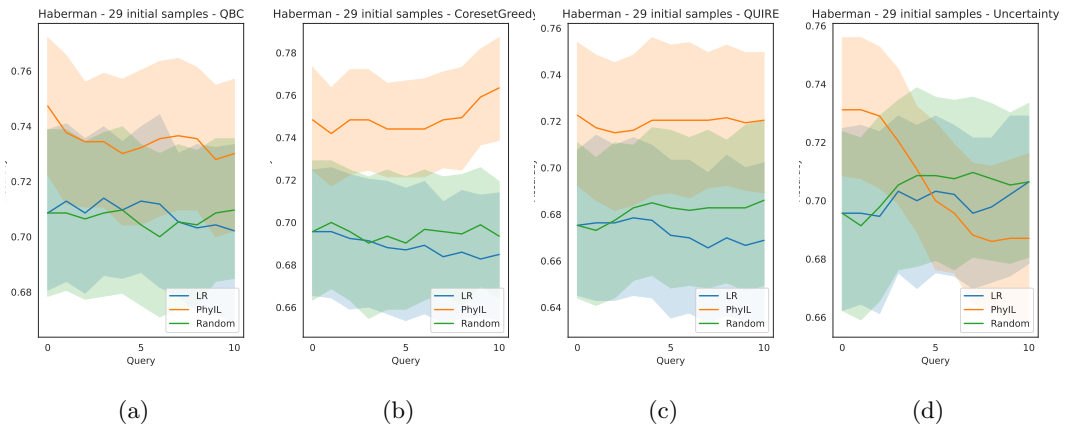
## 9.3 parkinson



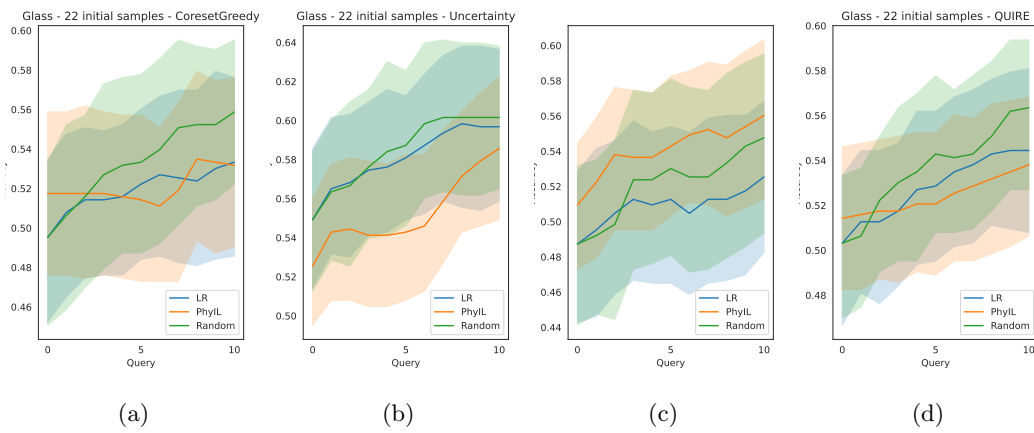
## 9.4 wine



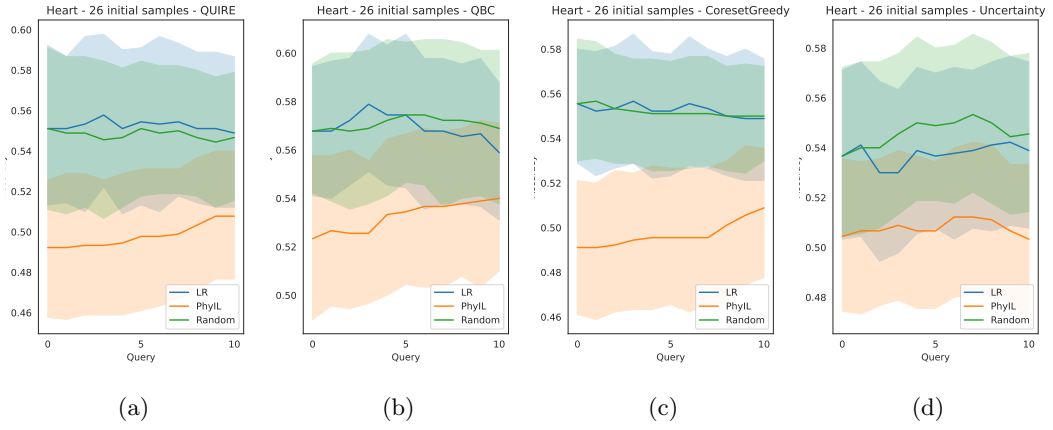
## 9.5 haberman



## 9.6 glass



## 9.7 heart



## 9.8 musk

