Comparison of Compositional Approaches to Taxonomic Binning for Metagenomics

Patrick Kimball, Ryan Nelson

02620 Machine Learning for Scientists – Group Project

May 5, 2021

Outline

- Introduction
- Methods
- Results
- Conclusions

Outline

- **Introduction**
- Methods
- Results
- Conclusions

Introduction

Metagenomics

- Analysis of genetic sequences discovered through broad-stroke sampling of an environment
- Goal is to reconstruct full genome of discovered organisms
- Taxonomic binning enables this: Classify samples into groups
  - Alignment strategy (BLAST)
  - Compositional strategy (ML classification)

Introduction

Motivation

- Previous approaches leverage discriminative or generative models
  - Vervier et al. Large-scale machine learning for metagenomics sequence classification. (2015)

- Our goal is to compare the performance of these strategies
- Our second goal is to develop an understanding of how to prepare metagenomics data for supervised machine learning

Introduction

Metagenomics Data

- Dataset comes from paper Vervier et al. (2015)
- For supervised learning, species label for each sequence is known
- Collected sequences vary significantly in length (1000 bp – 10M+ bp)
- Need a way to standardize the length for use in machine learning without losing information provided by sequences

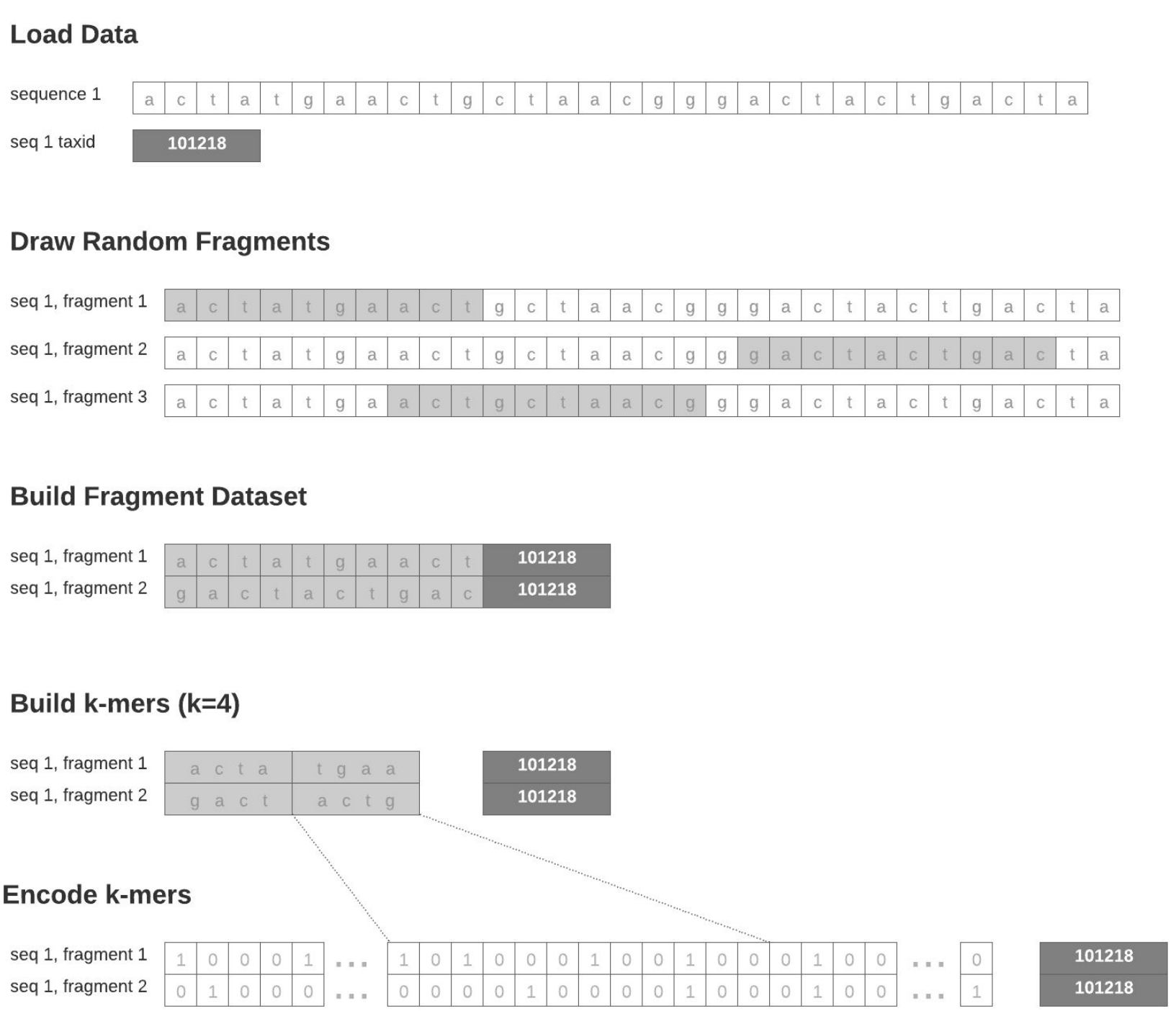
Outline

- **Introduction**
- **Methods**
- Results
- Conclusions

Methods

Dataset Prep

- Similar to paper
- Hyperparameters
  - Sample length
  - Coverage
  - K-mer size



Methods

Dataset Prep

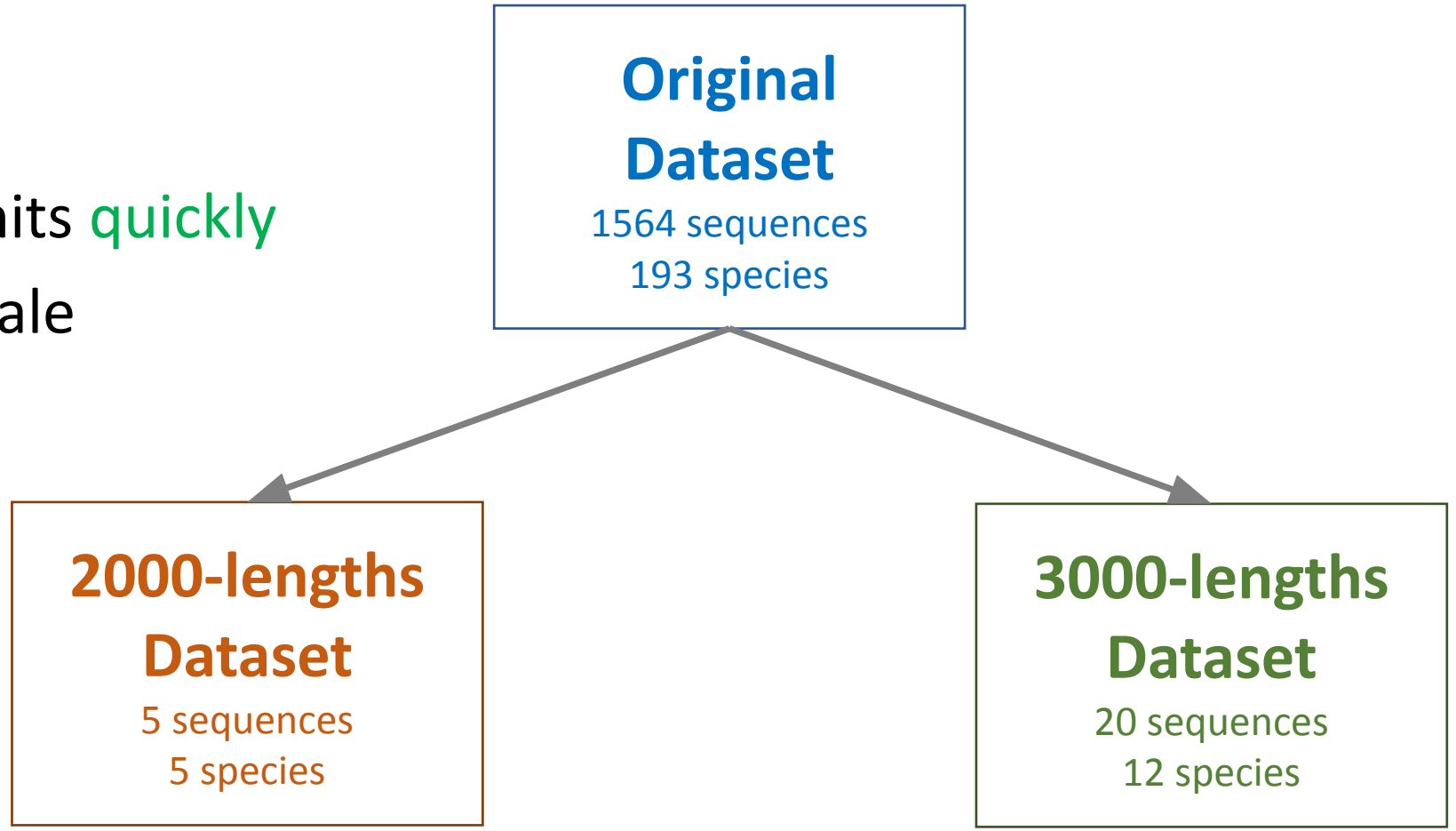
- Training set dimensions scale dramatically based on hyperparameters
- Explored sample length: 100 – 400, coverage: 0.1-400, k: 1 – 12

Dataset Dimensions	Sample Length	Coverage	K
314 x 400	100	1	1
122,706 x 369,078	100	400	10
:			
30,676 x 1,600	400	400	1
30,676 x 952,099	400	400	10

Methods

Dataset Prep

- Hit performance limits quickly
- Built two smaller-scale multiclass datasets



Methods

Model Selection

- Discriminative Models: SVM, Random Forest, Logistic Regression
- Generative Model: Naive Bayes
- We implemented Naive Bayes and (multiclass) Logistic Regression
  - *We used sklearn version of LR to explore advanced options.*

Note: SVM, Random Forest, Naive Bayes are “naturally” multiclass.

Methods

Scoring strategy

- Average recall
  - According to Vervier et. al: “We first compute the prediction recall within each species, i.e. the proportion of fragments originating from this species that are correctly classified and consider the average recall observed across species. **In a multiclass setting, this indicator is indeed less biased toward over-represented classes than the global rate of correct classification**.”

Outline

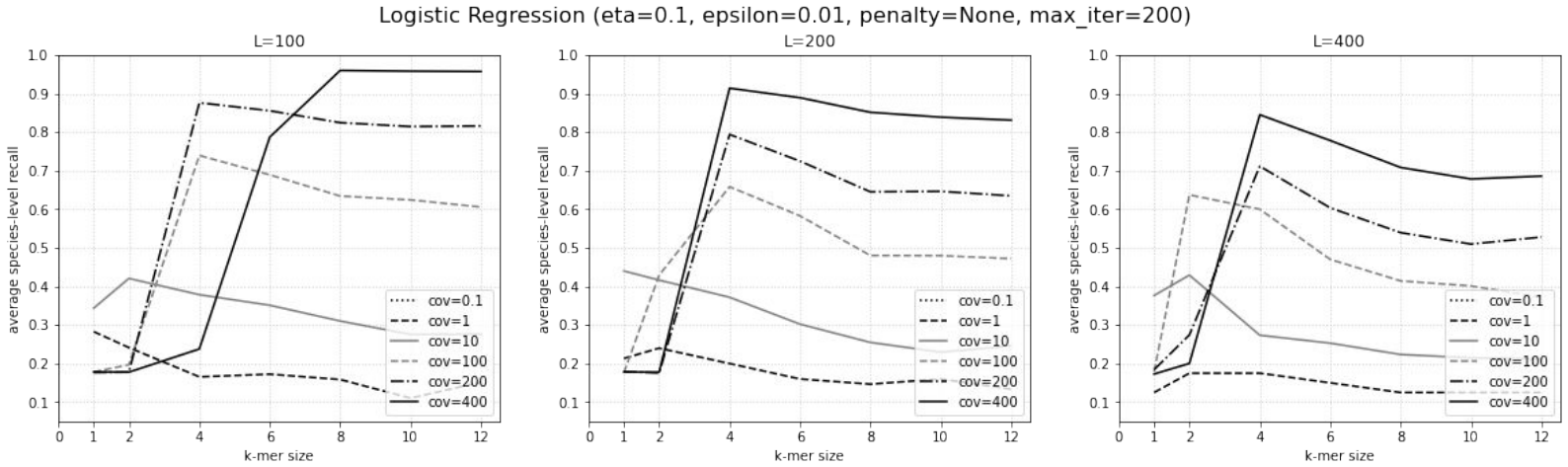
- **Introduction**
- **Methods**
- **Results**
- Conclusions

Results

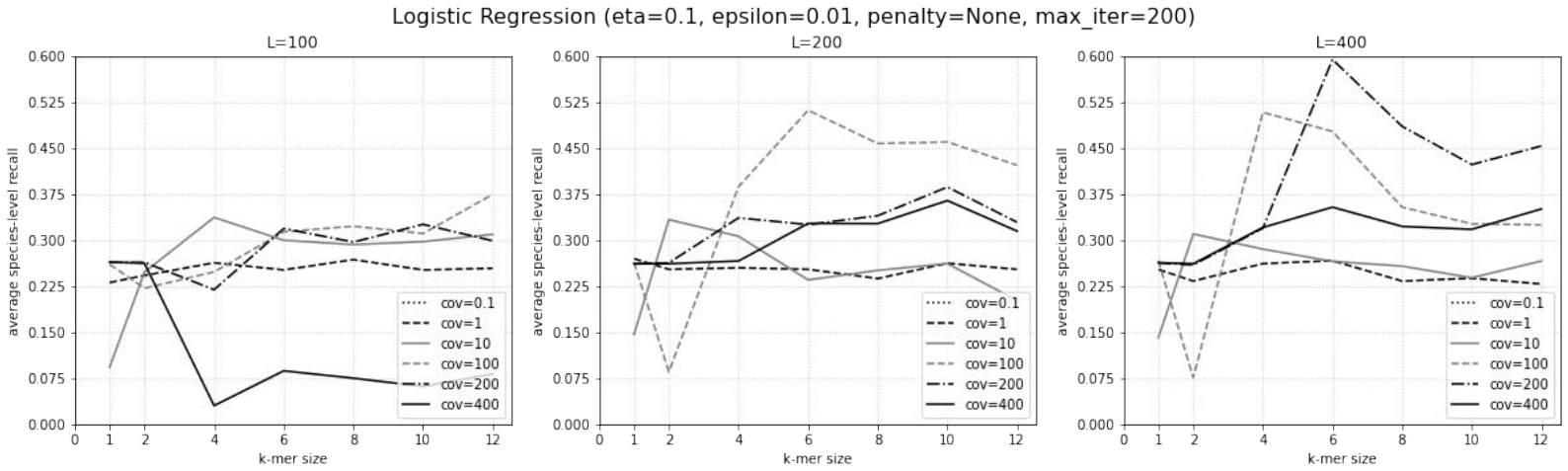
Discriminative Models

Logistic Regression
(our implementation)

2000-lengths Dataset



3000-lengths Dataset



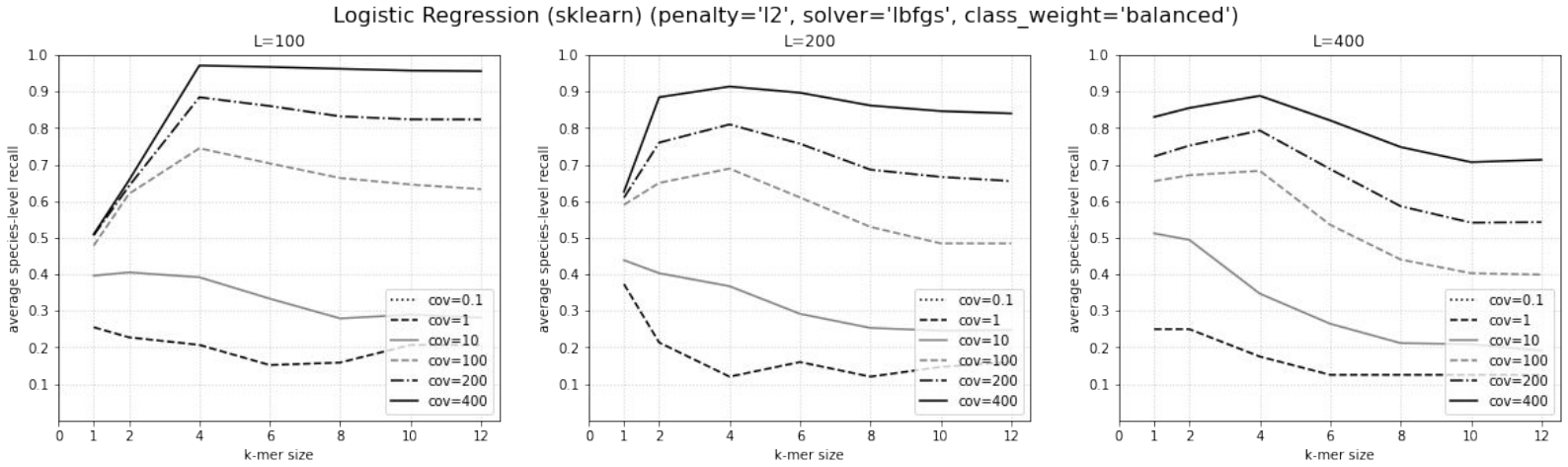
Results

Discriminative Models

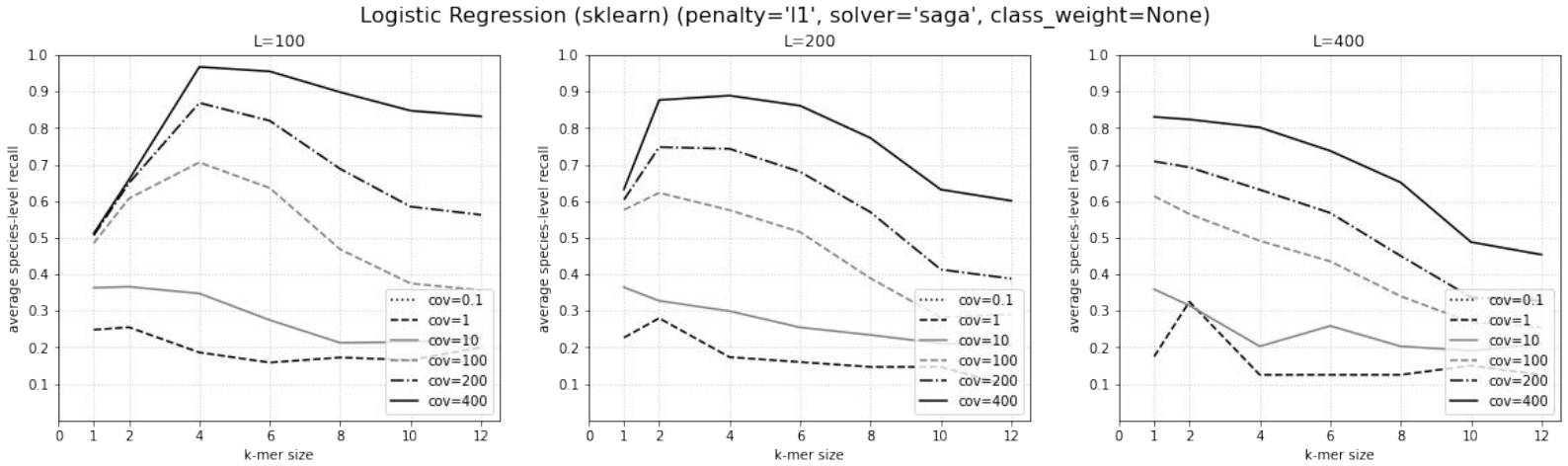
Logistic Regression (sklearn)

- Explored different penalties

2000-lengths Dataset L2 penalty, balanced weights



2000-lengths Dataset L1 penalty

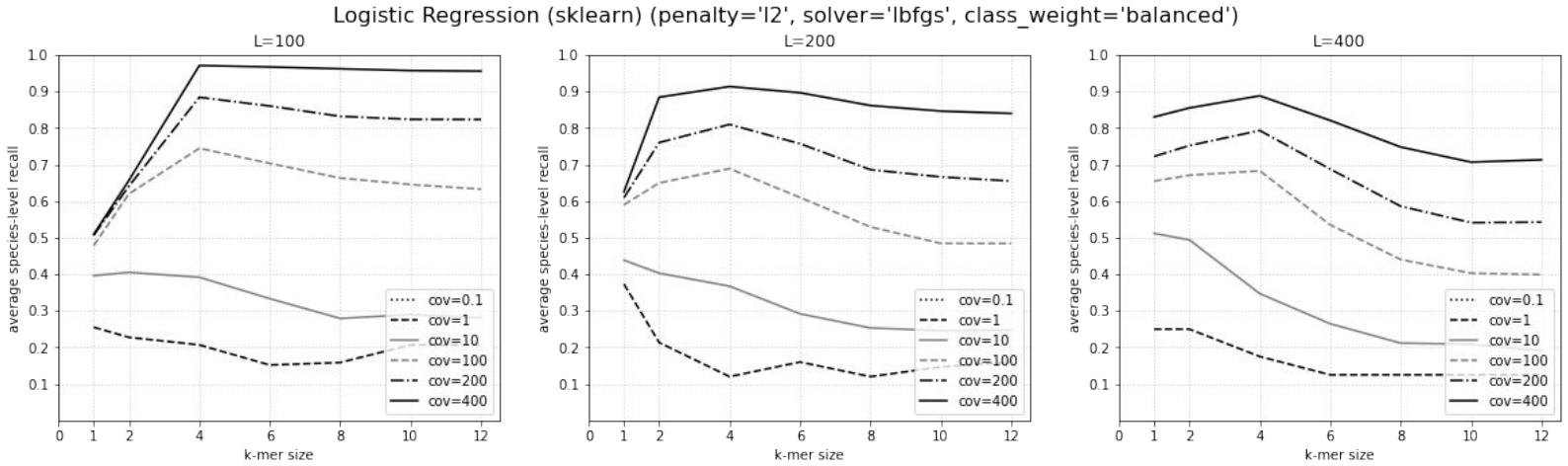


Results

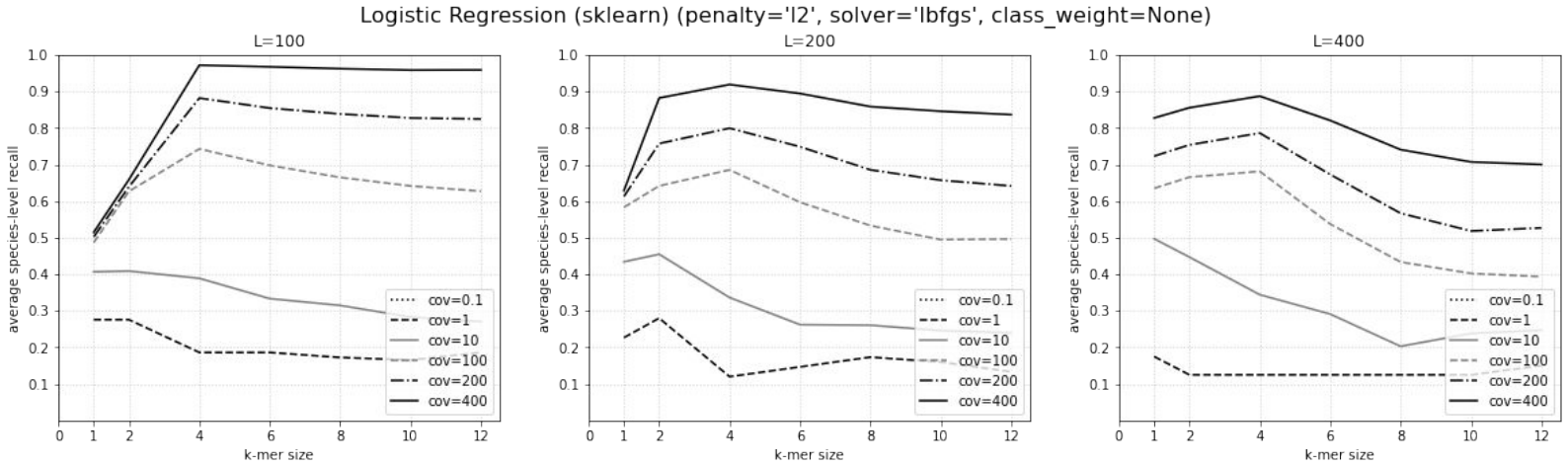
Discriminative Models

Logistic Regression (sklearn)

2000-lengths Dataset L2 penalty, balanced weights



2000-lengths Dataset L2 penalty

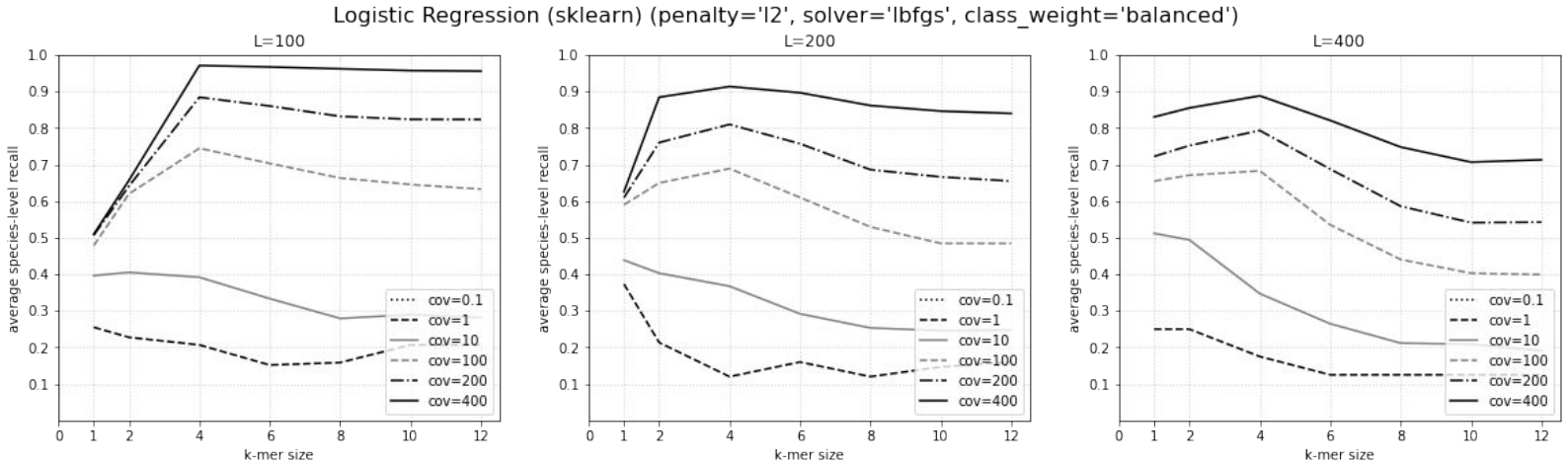


Results

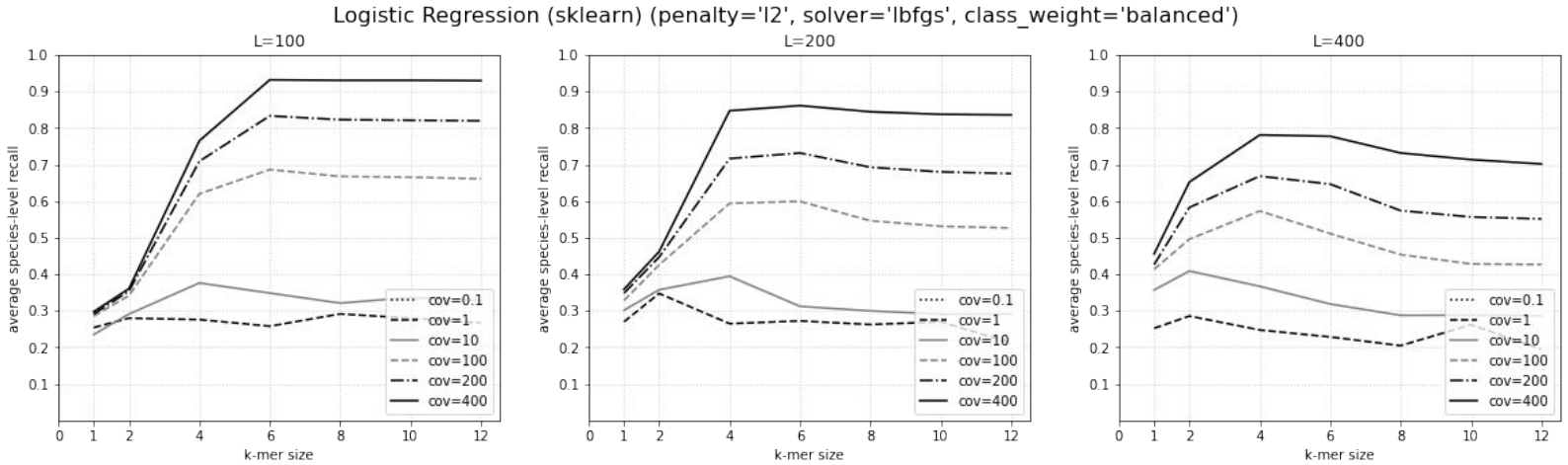
Discriminative Models

Logistic Regression (sklearn)

2000-lengths Dataset L2 penalty, balanced weights



3000-lengths Dataset L2 penalty, balanced weights

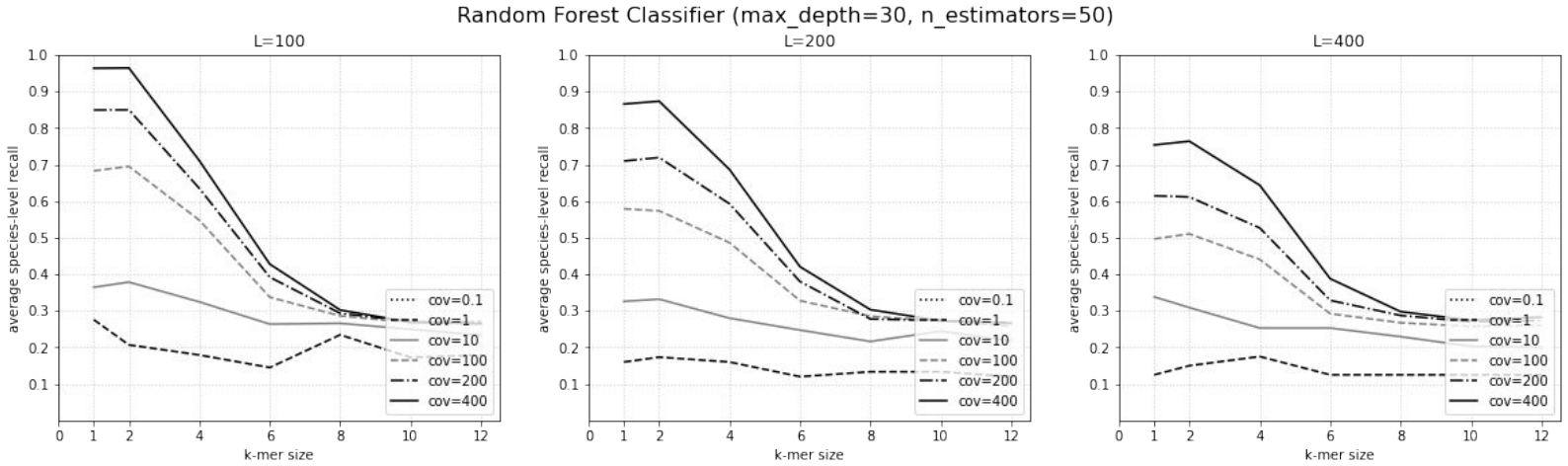


Results

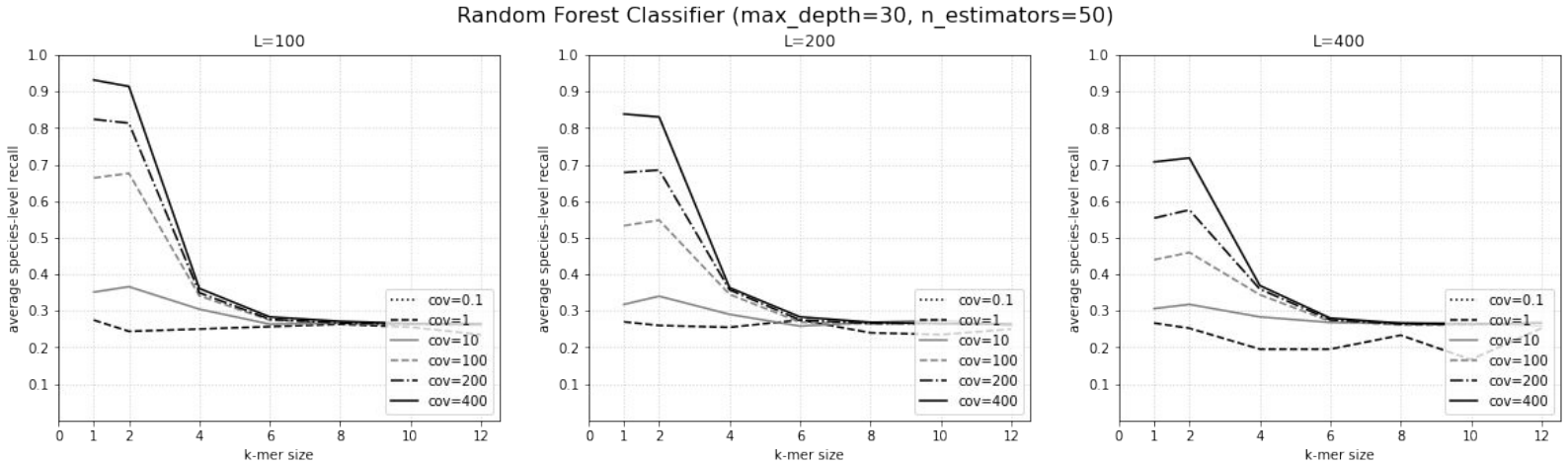
Discriminative Models

Random Forest

2000-lengths Dataset



3000-lengths Dataset



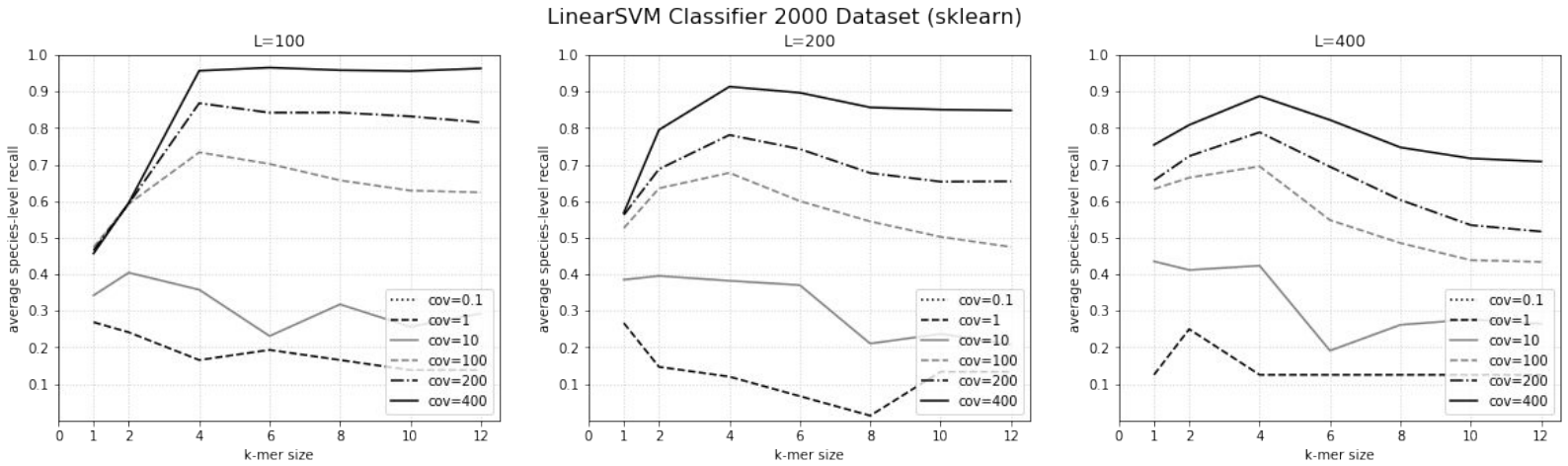
Results

Discriminative Models

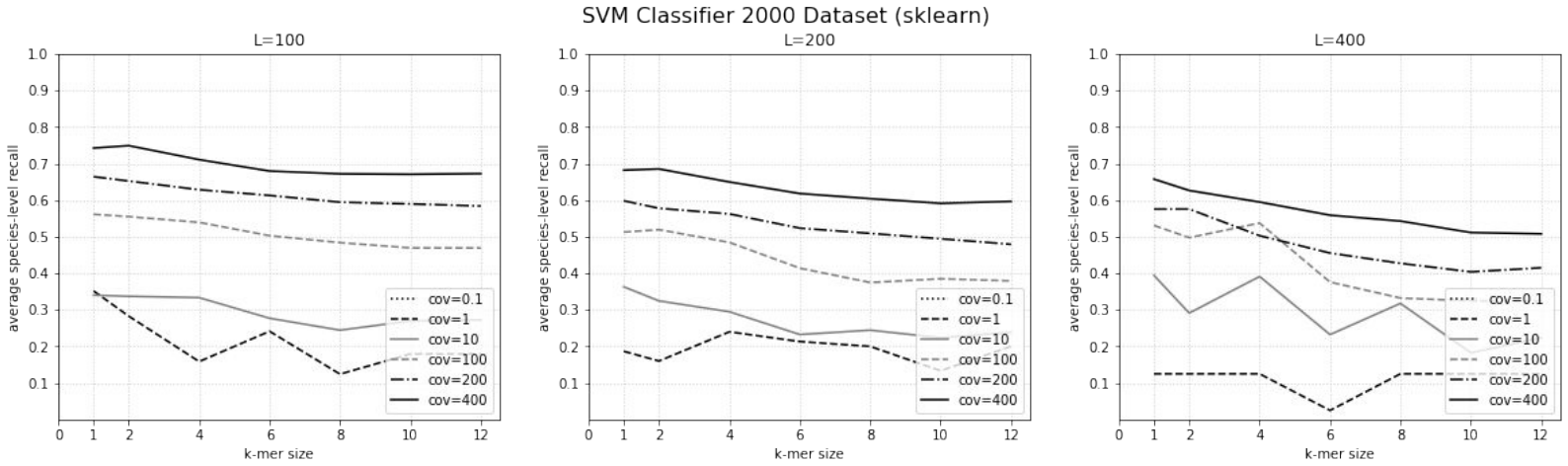
Linear SVM

- Explore different kernels

2000-lengths Dataset (linear kernel)



2000-lengths Dataset (RBF kernel)



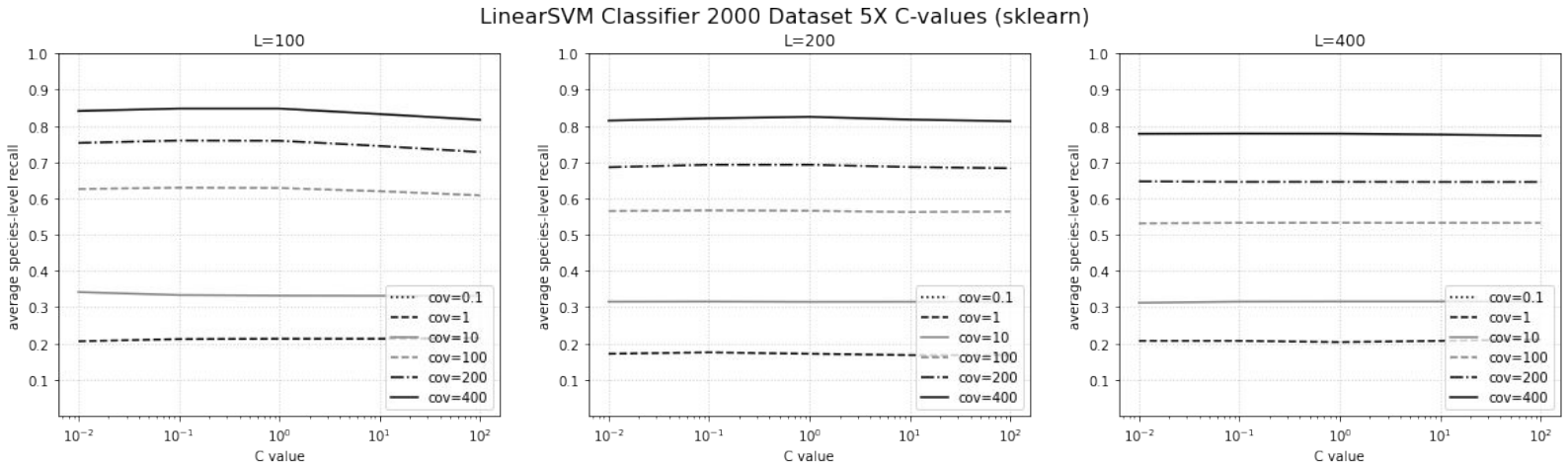
Results

Discriminative Models

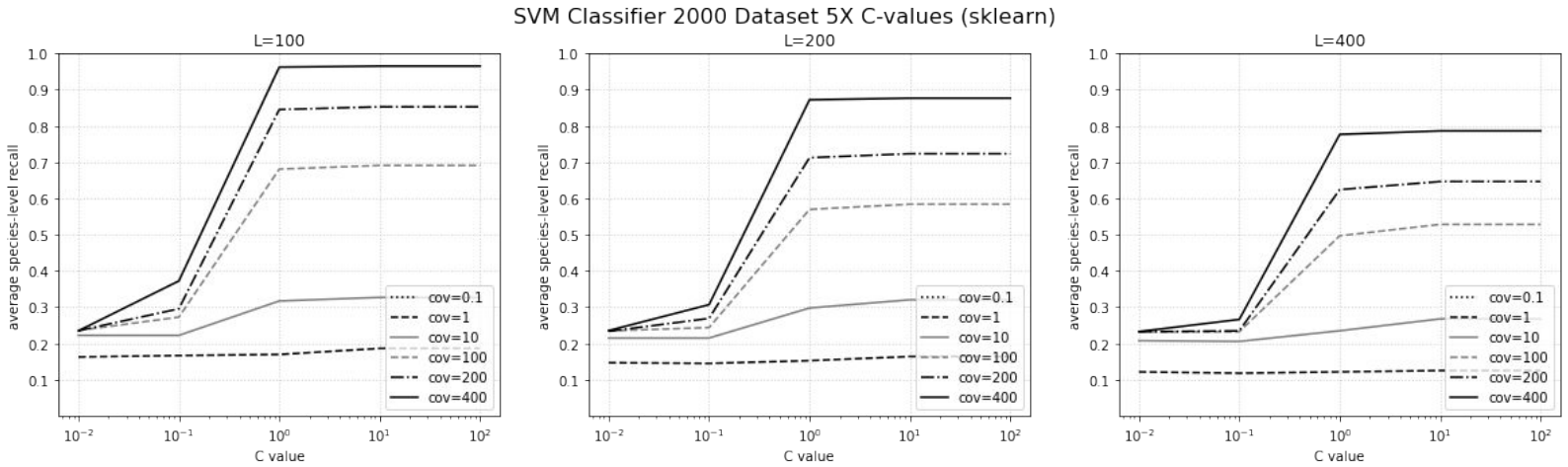
Linear SVM

- Explore different hyperparameters

2000-lengths Dataset (linear kernel, C-value)



2000-lengths Dataset (RBF kernel, C-value)

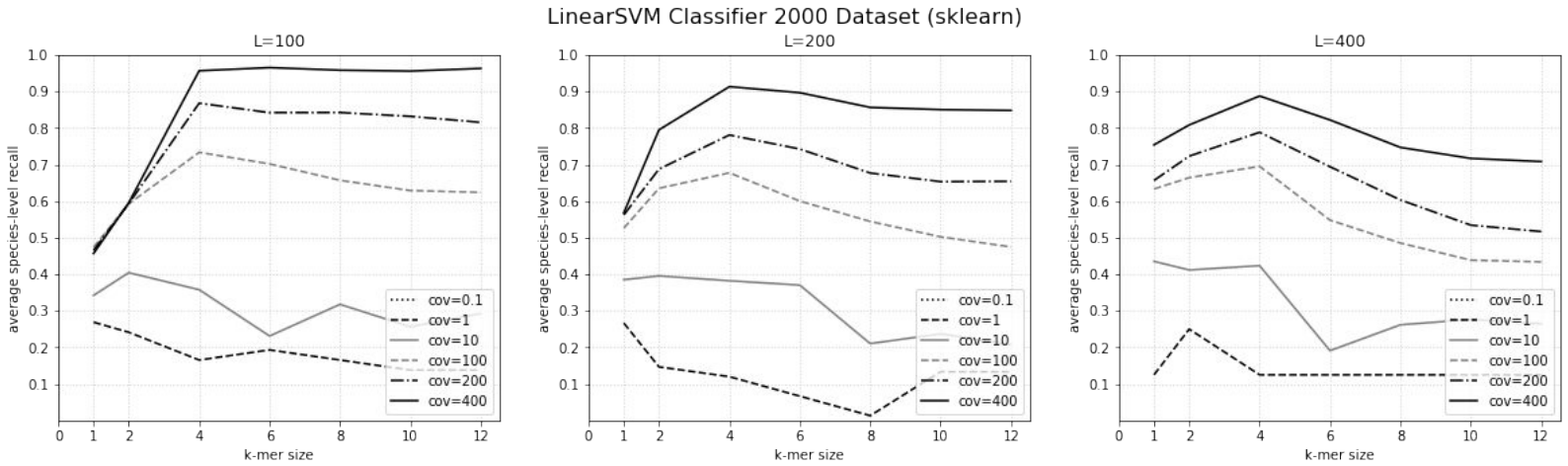


Results

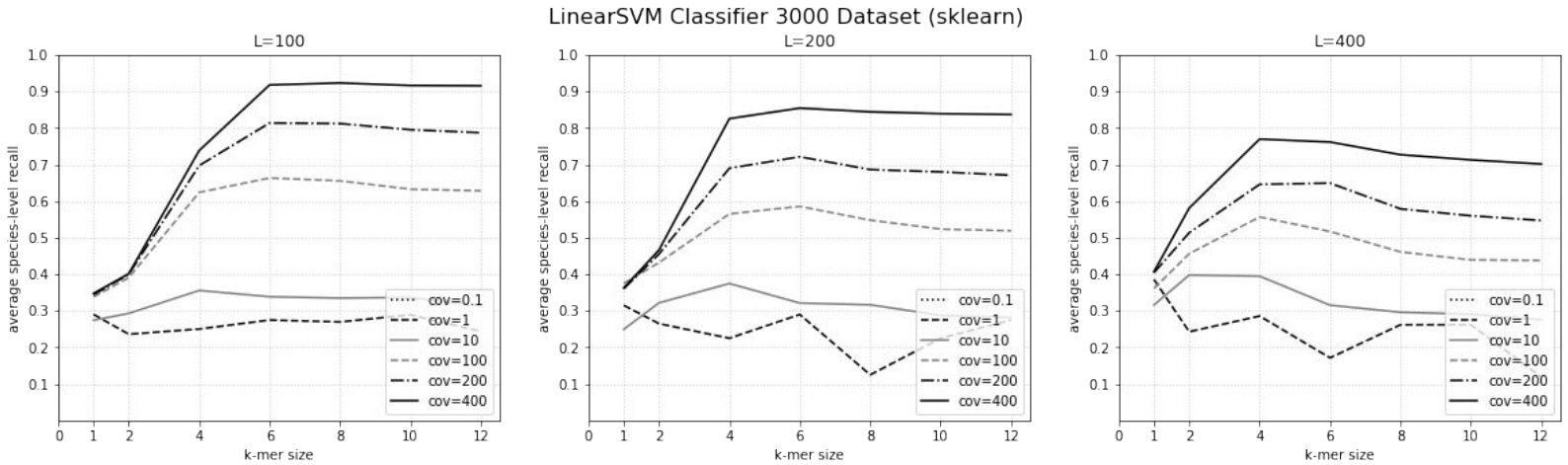
Discriminative Models

Linear SVM

2000-lengths Dataset (linear kernel)



3000-lengths Dataset (linear kernel)



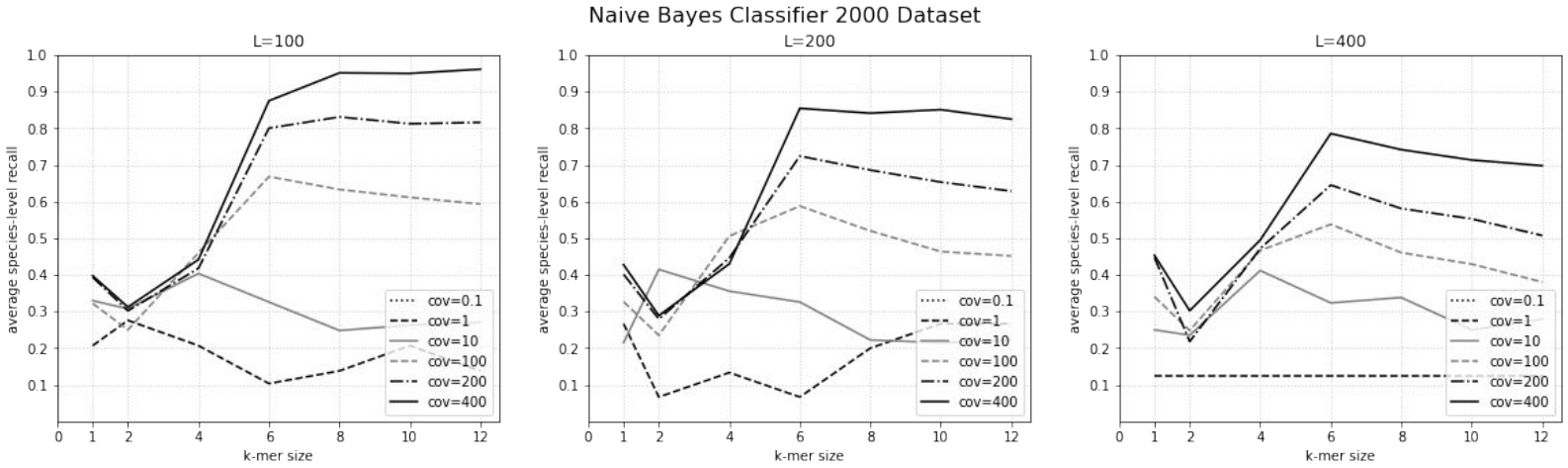
Note: RBF kernel runs very slowly for large datasets

Results

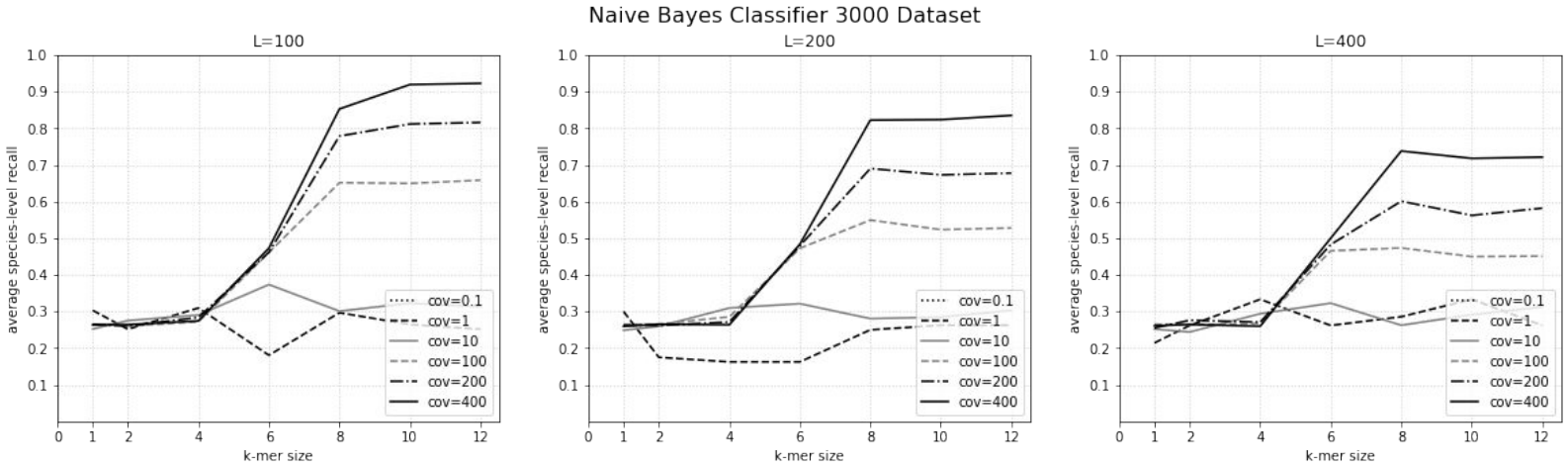
Generative Models

Naive Bayes

2000-lengths Dataset



3000-lengths Dataset



Outline

- **Introduction**
- **Methods**
- **Results**
- **Conclusions**

Conclusions

- **Fragment generation**
  - Current approach results in a very imbalanced dataset
  - Potential to bucket sequences based on length before generating fragments
- **Generative vs Discriminative models**
  - No clear difference between the two methods from our results
  - For all methods:
    - Increase sample length → lower average recall
    - Increase coverage → higher average recall
  - Random Forest was the only classifier that consistently performed better with smaller k-mer sizes
- **Metagenomics data is a challenge for standard ML packages!**

References

Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J. B., Vert, J. P. (2016). Large-scale machine learning for metagenomics sequence classification. Bioinformatics (Oxford, England), 32(7), 1023–1032. https://doi.org/10.1093/bioinformatics/btv683

Questions?