Identifying plasmids with machine learning (and deep learning)

Bill Andreopoulos

Jan Balewski

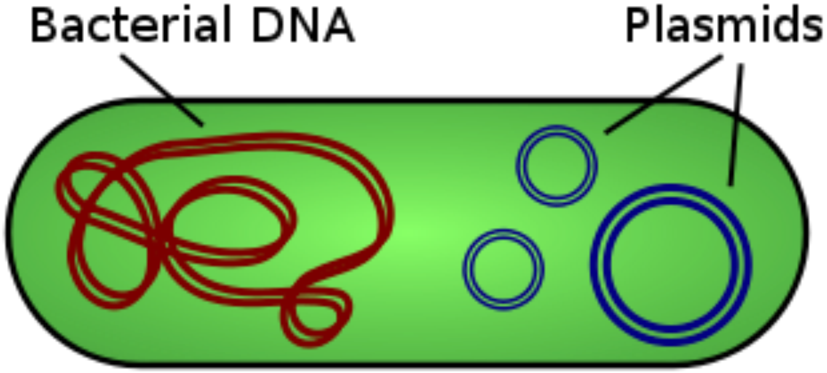
June 26, 2018

Outline

- Motivation
- Data set for training, data preprocessing
- Using features for classifying plasmids with ML
- Adding raw sequences into a Deep Learning tool
- Results:
  - Cross-validation on training set
  - Microbial genomics (IMG) dataset
  - MBARC-26 microbial mock community
- Production pipeline and deployment
- Conclusion and Future work

What are plasmids

“A genetic structure in a cell that can replicate independently of the chromosomes, typically a small circular DNA strand in the cytoplasm of a bacterium or protozoan. Plasmids provide a mechanism for horizontal gene transfer through conjugation.”



Identifying plasmids is hard: often plasmid sequences have become integrated in chromosomes, or vice versa.

Hypothesis: ML and Deep Learning can help predict plasmids.

Motivation for plasmid separation from genomes

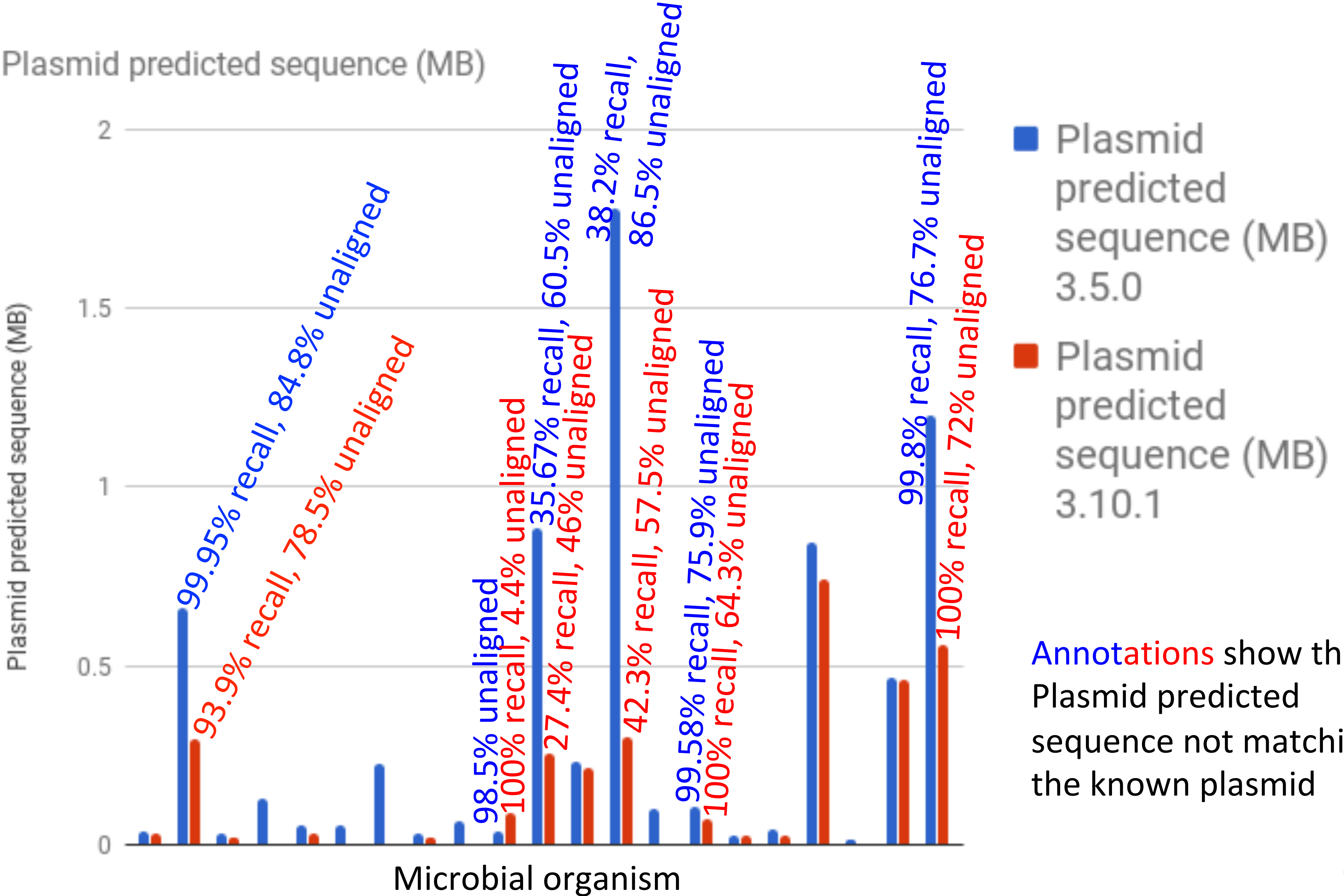
- Understand microbes in soil that play a role in biological nitrogen fixation
  - Microbes colonize plant root or have symbiotic relationship with plants
  - Plasmids with genes involved in nitrogen fixation are transferred via conjugation from soil microbes to root

2 approaches for plasmid separation from genomes

- Tried using *plasmidSPAdes* (*Bioinformatics*, 2016) to assemble plasmids from Illumina reads directly
  - Poor results for use in a production pipeline

- Decided instead to predict the plasmids post-assembly using 2 data types:
  - Extracted features and Raw sequence

2 approaches for plasmid separation from genomes



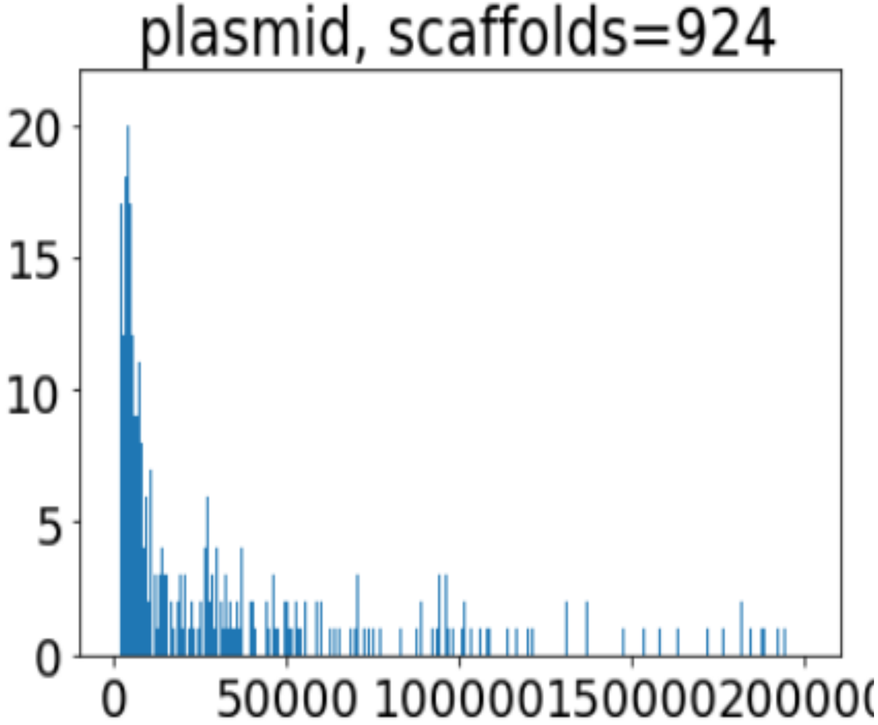
Annotations show the Plasmid predicted sequence not matching the known plasmid

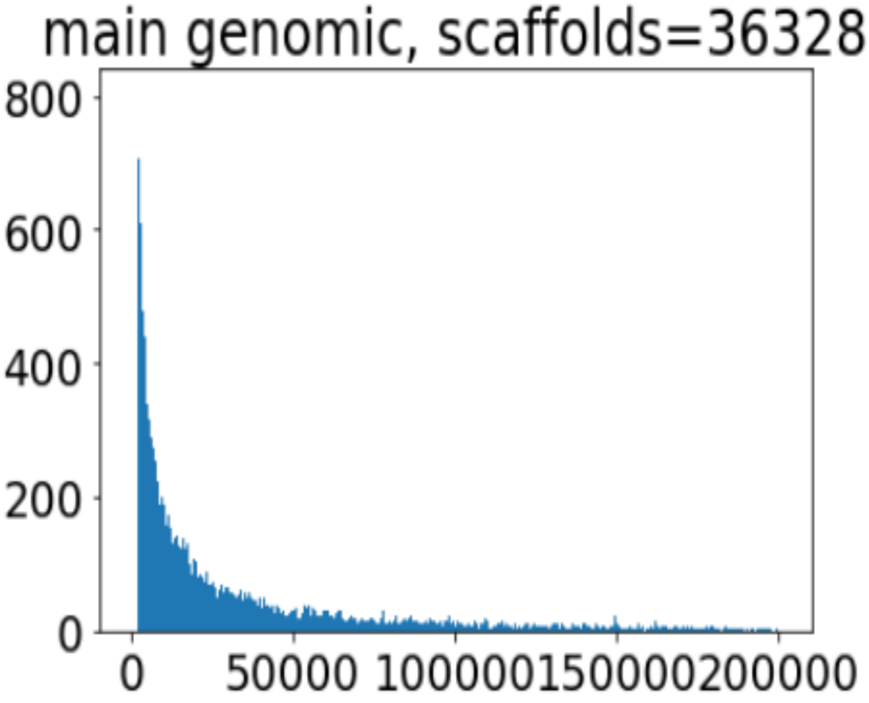
Datasets

Plasmids: Used ACLAME plasmids dataset with 1095 scaffolds because it is manually curated. Refseq.plasmids has many errors

Microbial genomic: Refseq.microbial - removed plasmid and mito sequences, then subsampled 40k seqs

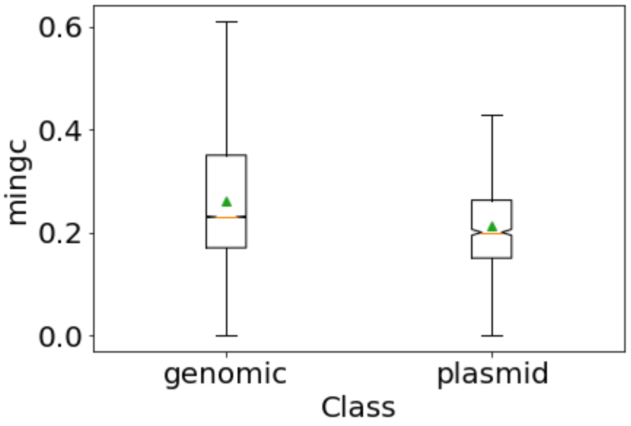
Kept just scaffolds of sizes 2KB-200KB. Length histograms:



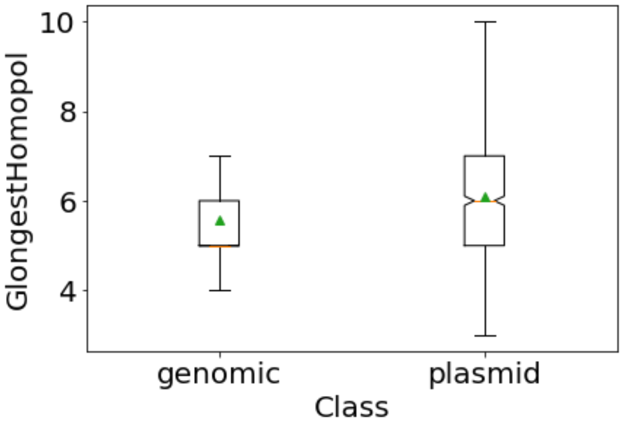


Features extracted from scaffolds

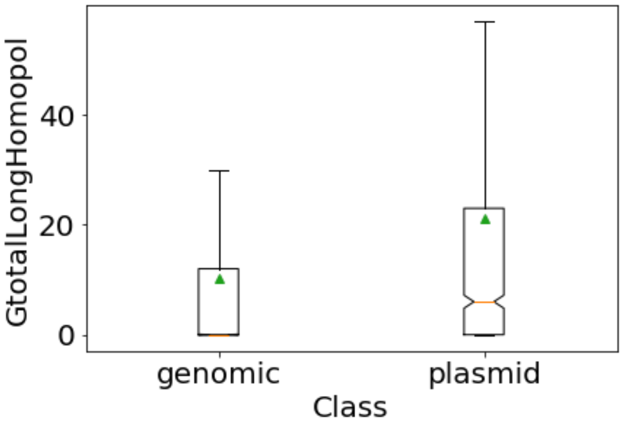
Min GC content in win-100b



G longest homopolymer



G total homopolymer len>5



Other features with chi2 p-values <0.01: GC content overall, MaxGC in windows of 100b, A/C/G/T/* longest homopolymer sequences, A/C/G/T/* total homopolymers len>5, sequence lengths.

Classic ML tools, do feature vectors have predictive power?

Trained and validated just on feature vectors (no raw sequences)

Cross-Validation in scikit-learn: 20 random shufflings with 20% used as test data. Mean ROC-AUC:

Logistic Regression 79.6%

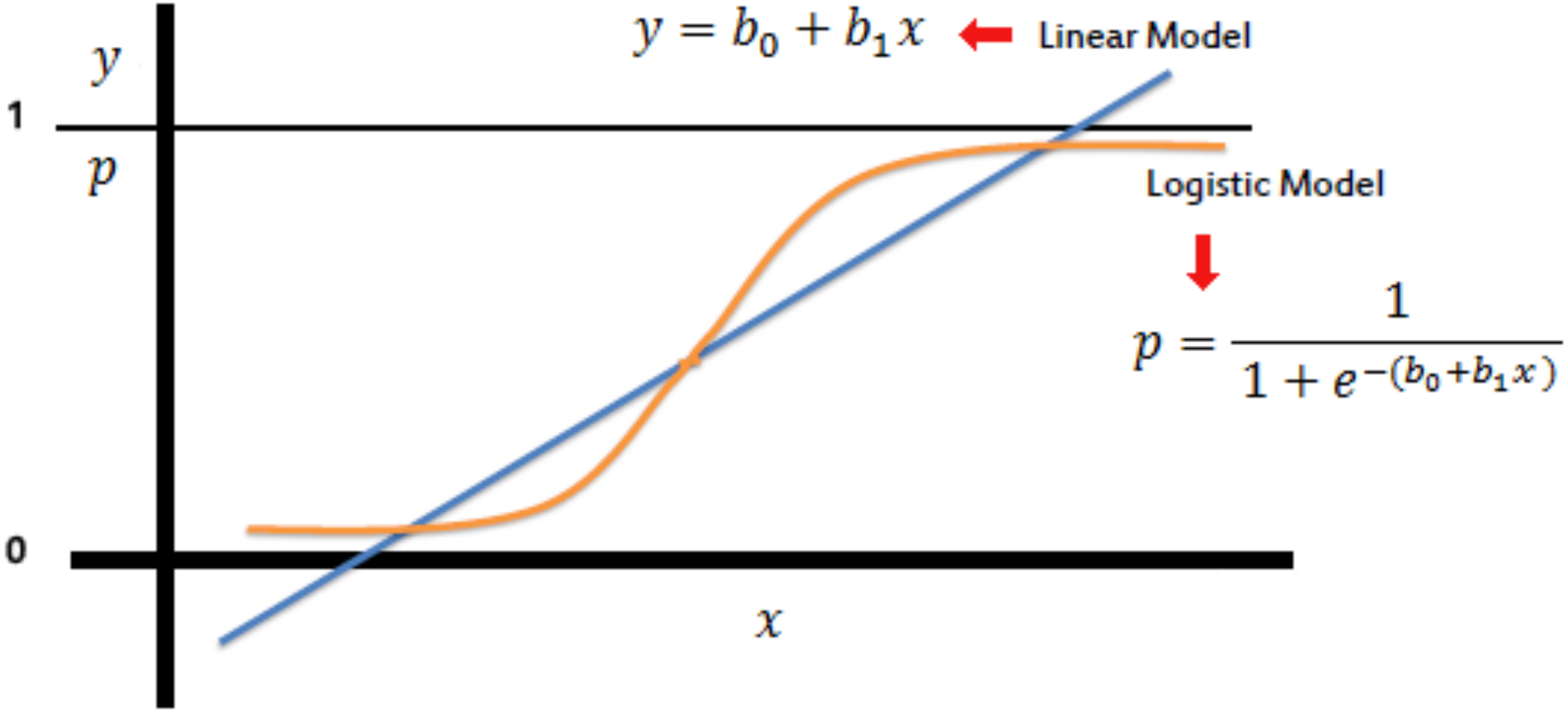
GaussianNB 70.35%

DecisionTree 77.7%

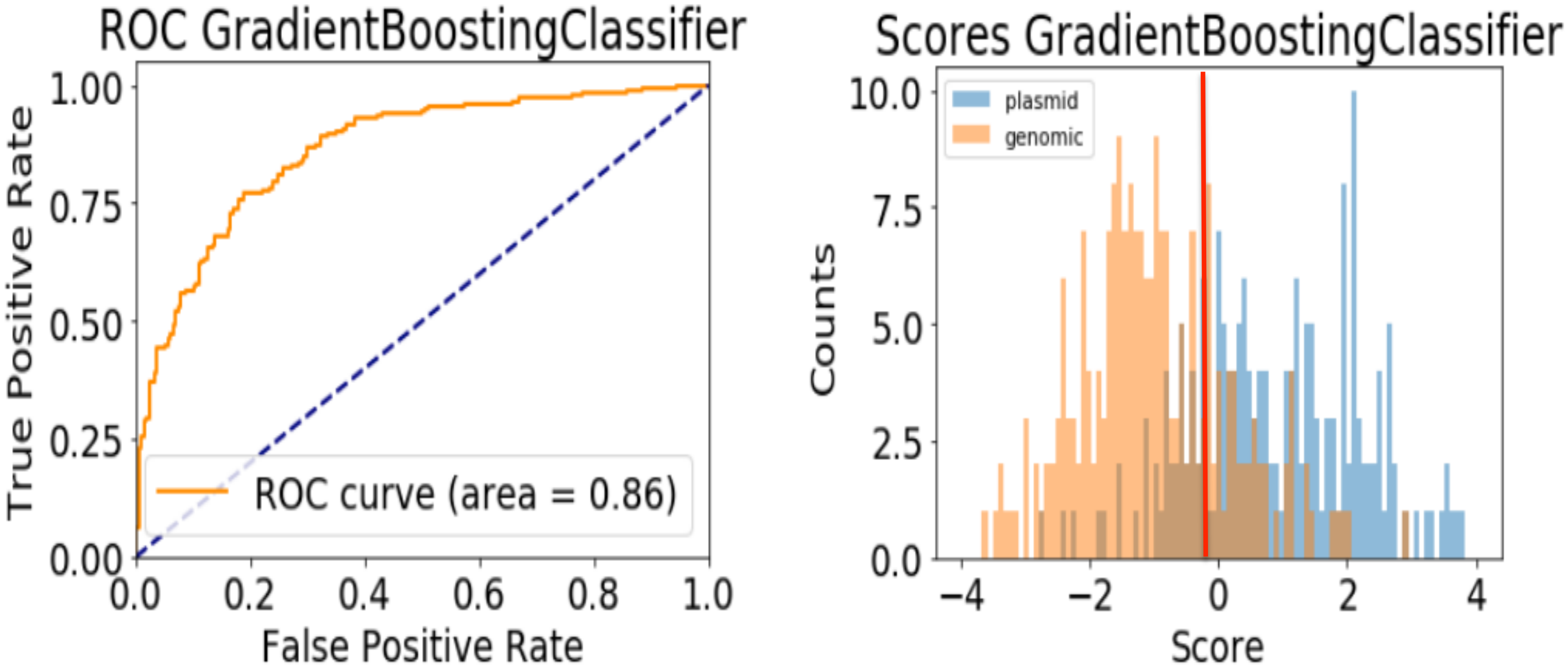
SVM 66.5%

Gradient Boosting Classifier 86.9%

Logistic regression (79.6%)

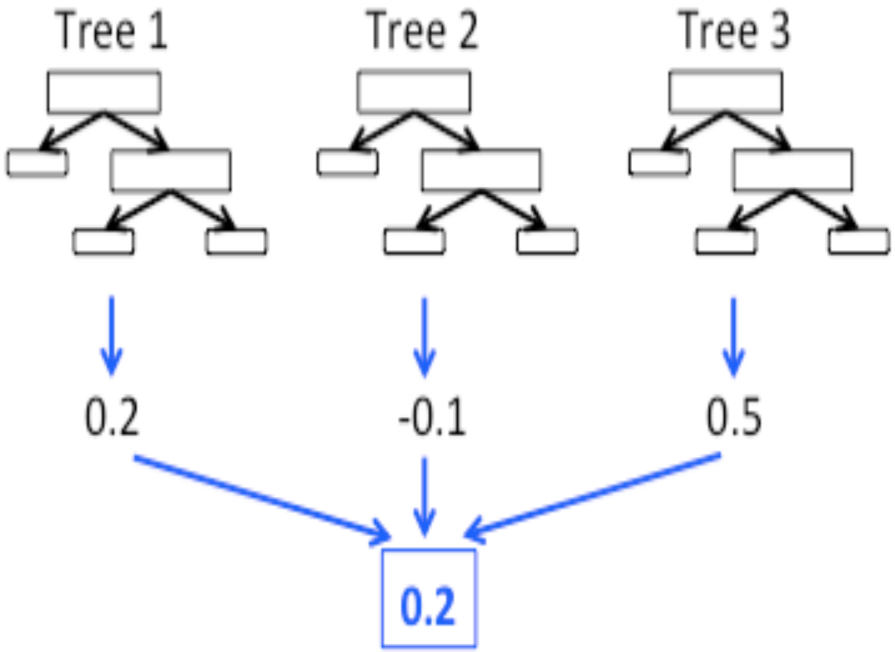


Gradient Boosting Classifier on 20% test set and 20 validation splits

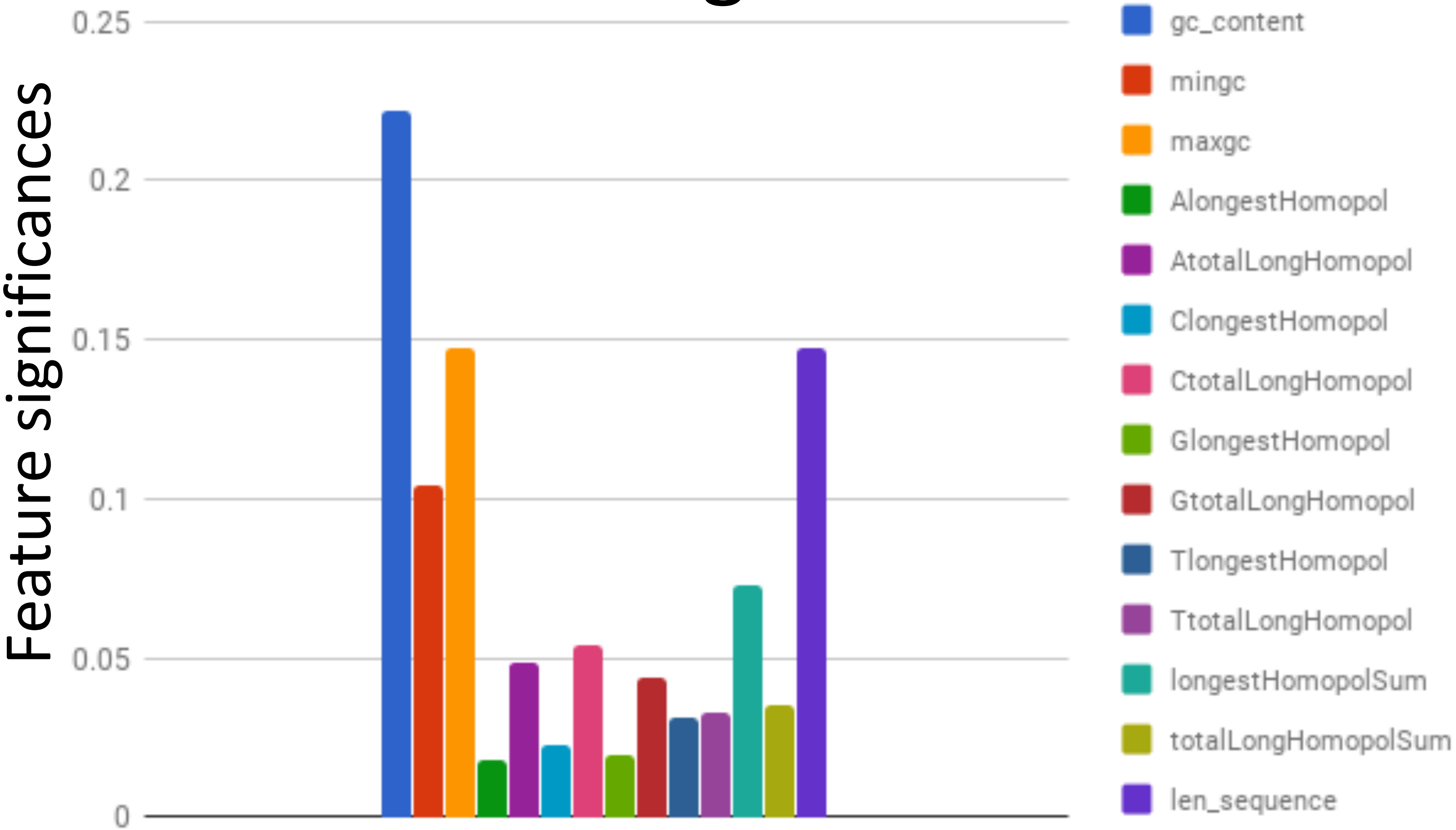


Gradient Boosting Classifier (86.9%)

Gradient boosting builds an ensemble of trees one-by-one, then the predictions of the individual trees are summed



Gradient Boosting Classifier Feature significances



Other possible features to use in training

Most frequent dimers…heptamers in a scaffold

370 COG gene models that are genome specific

- Too time consuming to train on because requires all COG genes to be input for probability computation –> high runtime computing COG hits
- Input vector is long

Problem of very unbalanced classes

- ACLAME plasmid dataset << refseq.microbial
- Initially tried to upsample the smaller
  - Upsampling the smaller dataset resulted in overfitting since many sequences were repeated many times.
- Downsampled refseq.microbial instead.

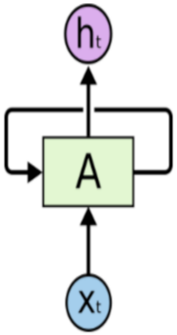
Next

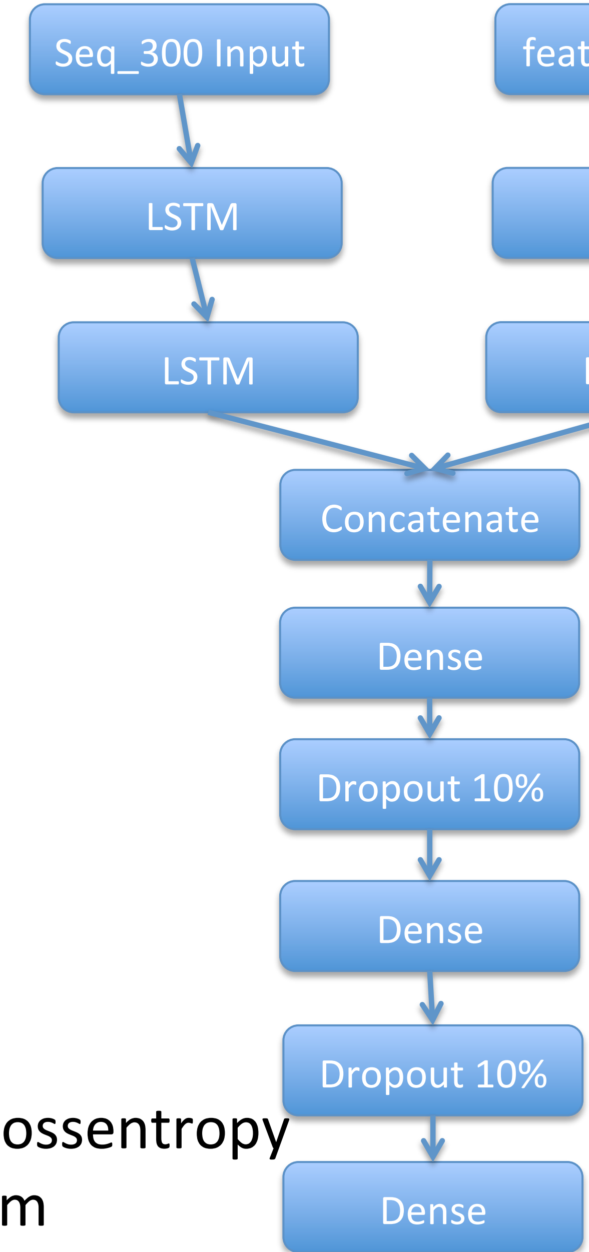
- Include the raw sequences in training with deep learning

Automated plasmid finder with Deep Learning (Keras)

- Training input: for each scaffold (of arbitrary length)
  - Draw 50-100 300bp samples from 1 scaffold (longer scaffolds contribute more samples, but don’t overwhelm)
  - Train on each sample
- Prediction for 1 scaffold:
  - Trained model outputs score for each 300bp sample: [0,1]
  - Compute average score +/- stdev over all samples
  - if [avg score >0.5+2*stdev] → plasmid
  - else if [avg score <0.5-2*stdev] → non-plasmid
  - else → ambiguous

Automated plasmid finder with DL (Keras)




Seq_300 Input

features Input

LSTM

Dense

LSTM

Dense

Concatenate

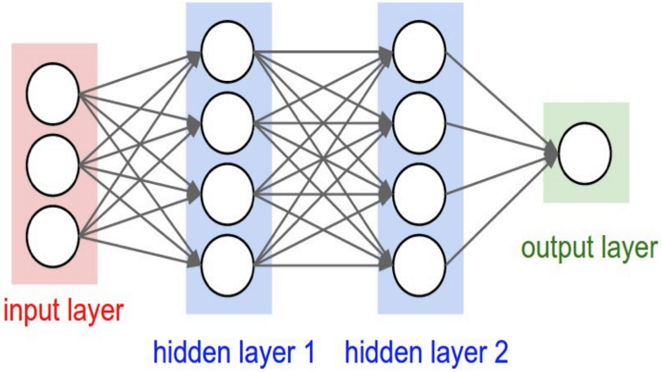
Dense

Dropout 10%

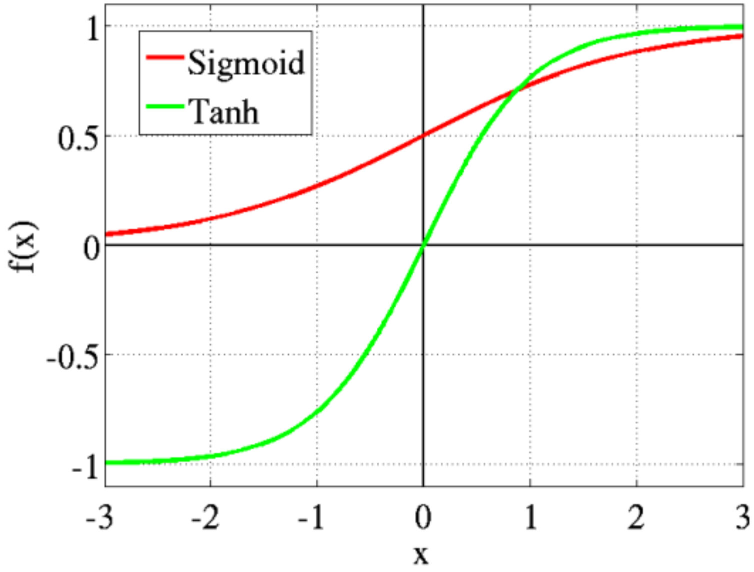
Dense

Dropout 10%

Dense



activation=tanh, relu, sigmoid



Output is [0-1]

Loss=binary_crossentropy

Optimizer=adam

DL - Cross-validation method

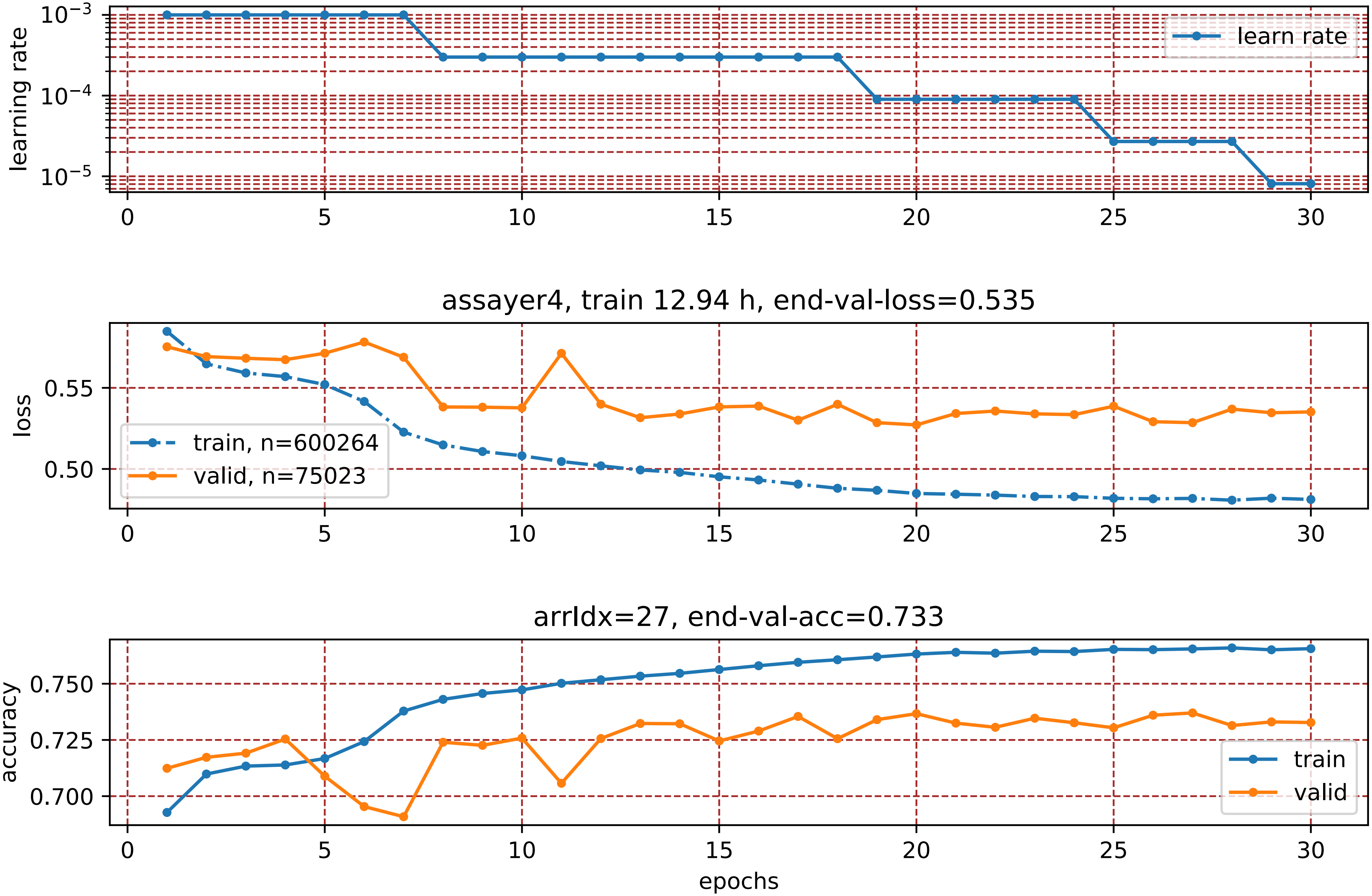
Split into 6 segments, with 1 segment test

5-fold cross validation with 1 out of 5 validation

For each segment the model is saved as an h5 file

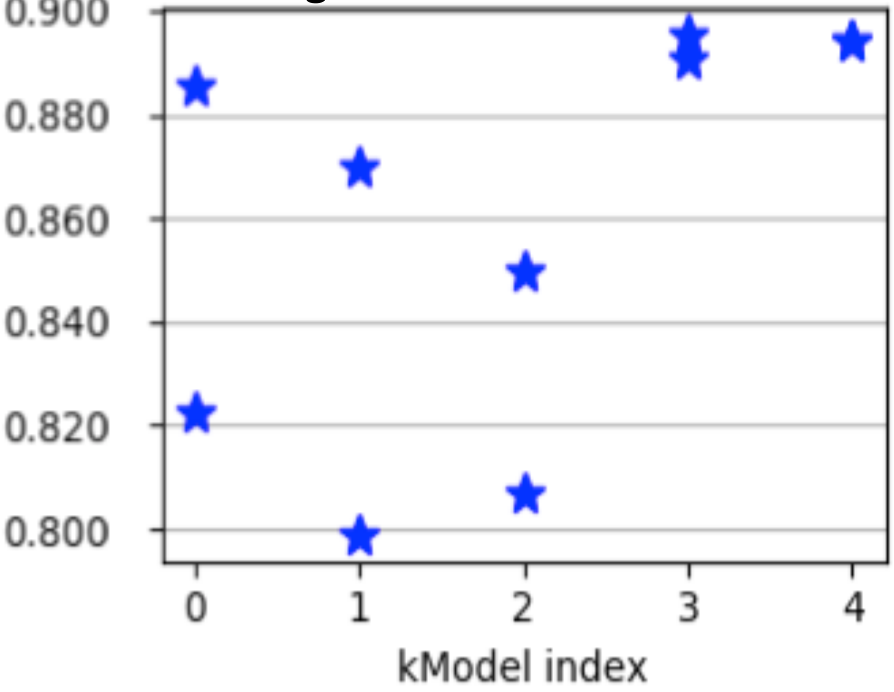
Training fold	seg0	seg1	seg2	seg3	seg4	seg5 - Test
1	Train	Val	Train	Train	Train	Test
2	Train	Train	Val	Train	Train	Test
3	Train	Train	Train	Val	Train	Test
4	Train	Train	Train	Train	Val	Test
5	Val	Train	Train	Train	Train	Test

DL - Results on fold ⅗ for the validation data (ACLAME+refseq.microb)

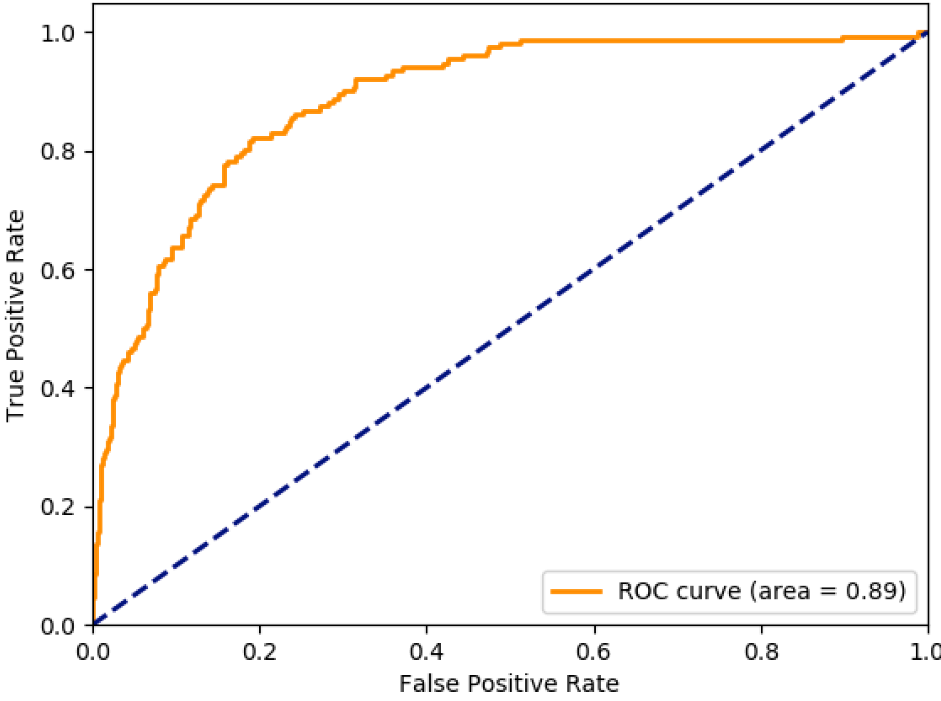


Deep Learning results on test data from ACLAME+refseq.microbial

Avg ROC-AUC=0.878



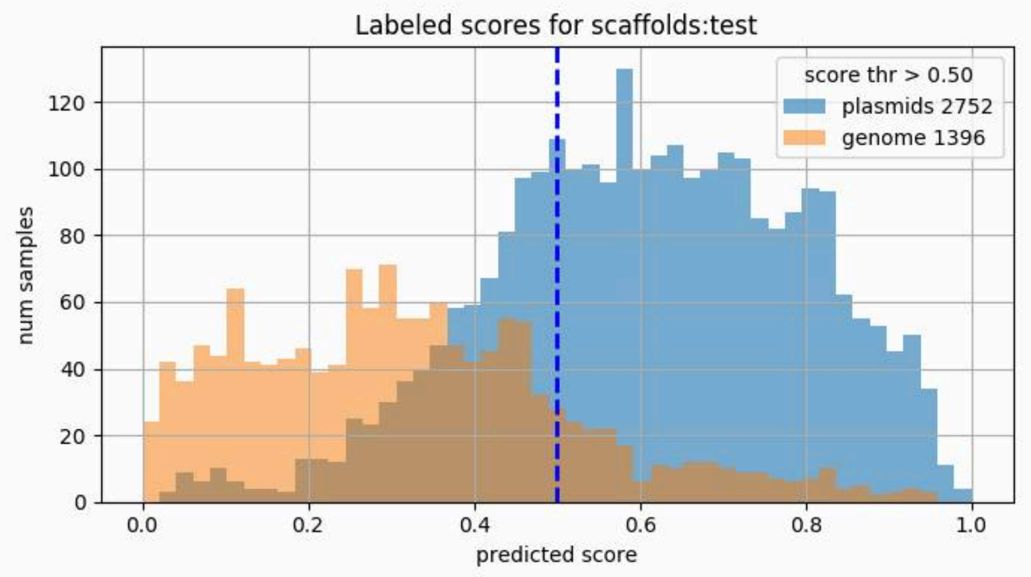
ROC-AUC=0.89



Test dataset from IMG

- Downloaded from IMG all 1834 organisms with at least one plasmid
- 6820 scaffolds in total: 3093 plasmids + 3727 genomic
- Used only scaffolds of length 2k-200k bases

	Classified plasmid	Classified genomic	
True plasmid	**2064**	**685**	**2749**
True genome	206	1136	1342
	2270	1821	4091



Precision ~91% (TP/TP+FP rate = 2064/2270 = 90.9%).
Recall is 75% (TP/TP+FN = 2064/2749 = 75%).
264 ambiguous predictions, evenly split between genome/plasmid

>90% of what is predicted as plasmid is true plasmid

MBARC-26 mock community*

- 26 microbial organisms, 38 scaffolds.
  - 13 scaffolds in 7 organisms are plasmids

	cBar (Bioinformatics, 2010)	Naïve Bayes	Deep Learning
TP	7	6	9
FP	4	0	0
FN	5	6	3
Precision	63.6% (7/11)	100%	100%
Recall	58.3% (7/12)	50%	75% (9/12)

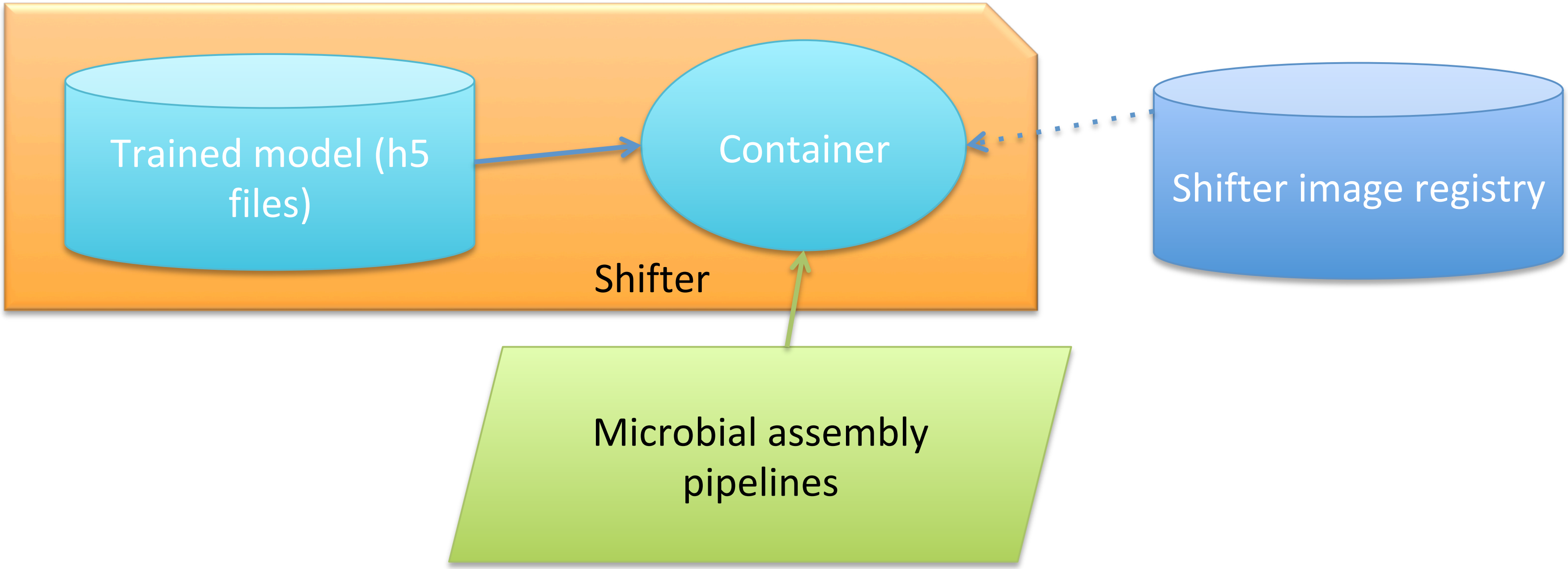
*E Singer, B Andreopoulos et al. Next Generation Sequencing Data of a Defined Microbial Mock Community. Scientific Data 3, 160081, 2016.

Production pipeline

- Requirements:
  - Scalability to ~500 microbial assemblies weekly
  - Input is assembled fasta
  - Output is a CSV file: per-scaffold classification of MAIN, PLASM, AMBIG, along with a score representing confidence
  - Reusability, reproducibility : Docker containerization

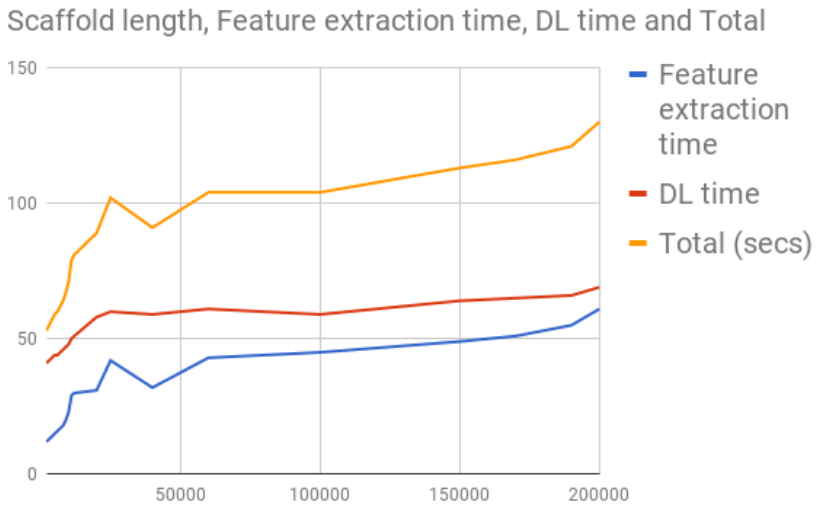
Docker image

- Need Docker image with Keras dependencies
- DL Model is stored as .h5 files, added into the container



Production pipeline

- Runtimes on Cori
  - Training runtime is ~12.94 hours for the ACLAME+refseq.microbial dataset with 41K sequences
    - 30 epochs - 26 minutes per epoch
    - Used 5 Intel Xeon “Haswell” nodes with 120GB, 16 cores
  - Prediction runtime is <2 minutes per scaffold on a single node



Conclusion

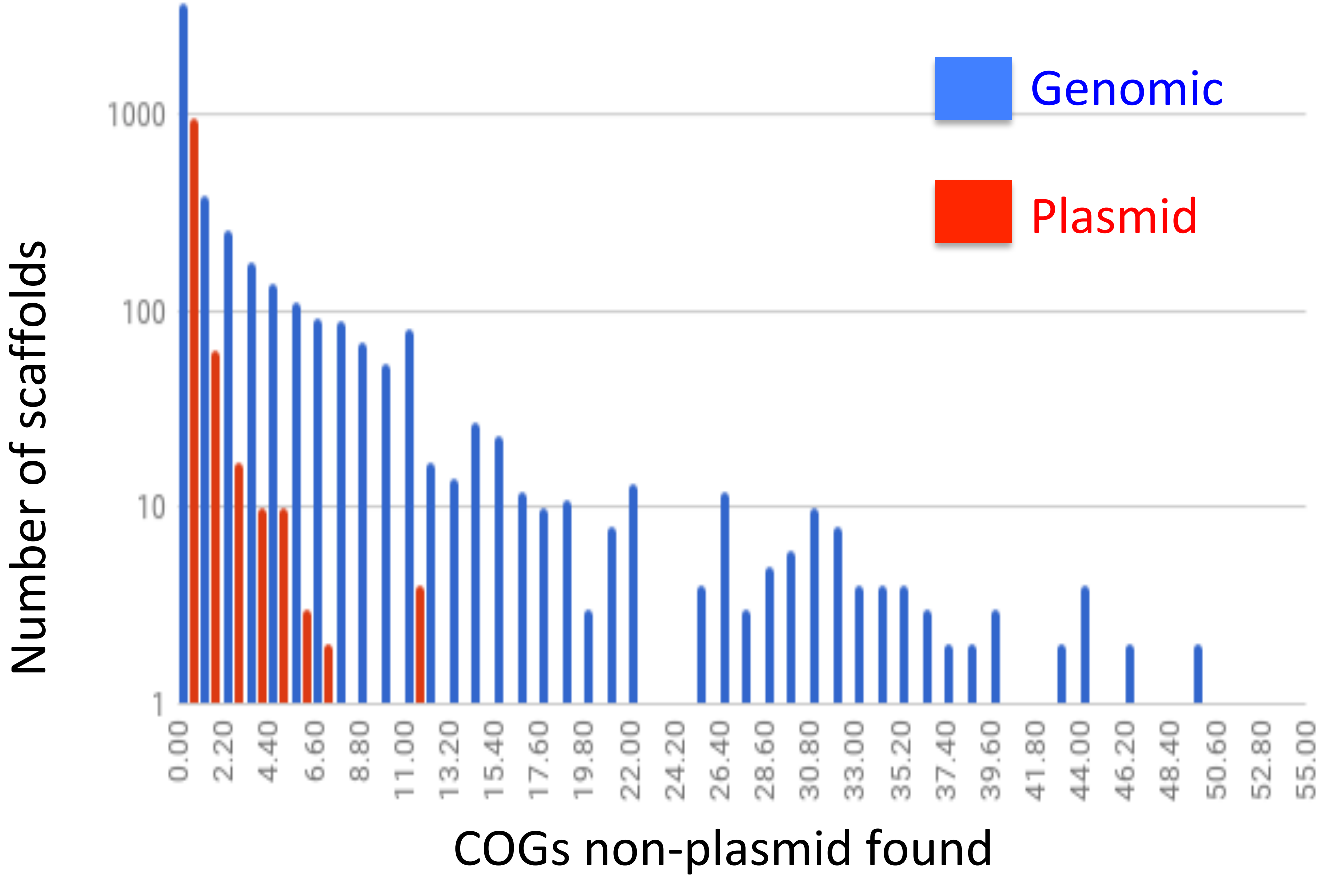
- Prediction of plasmids is complicated in large genomic datasets, complex feature relationships
- It is possible to find plasmids with high precision with ML
- Can further train on specific plasmid examples to improve recall
- If the dataset is highly unbalanced, small error rate will amplify

Future work

- Retrain with plasmids that were misclassified to improve recall

- Science: Do P-value study to find genes transmitted between plant-microbes:
  - Find genes enriched in:
    - Plasmids vs. genomes
    - Plant-associated vs. non-plant-associated microbes.
    - Root associated vs. soil associated microbes.
  - These could be symbion-genes that are important for biological nitrogen fixation or pathogenic resistance genes

Other features: 370 non-plasmid (genome-specific) COGs



Conclusion: the genomic COGs are more frequent in the genome sequences than in the plasmid (ACLAME) sequences.