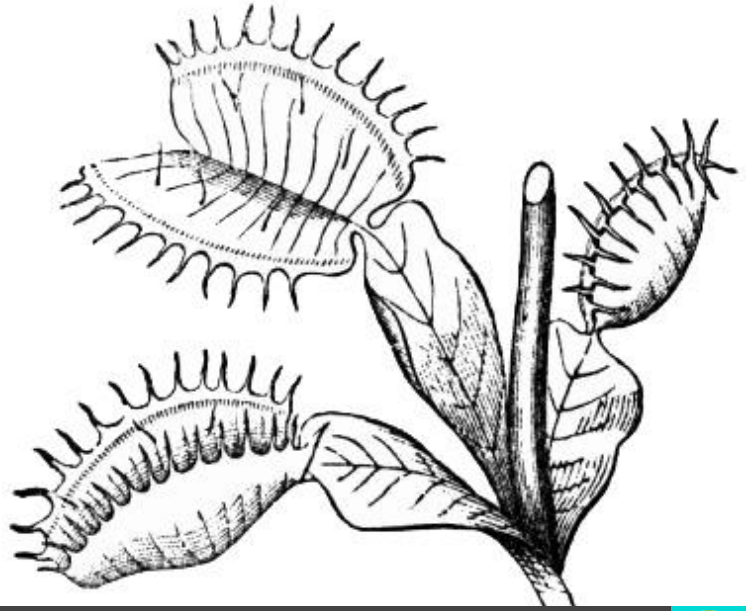
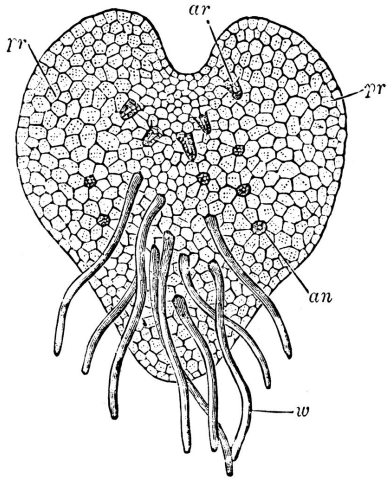
Biollante

DSI

Paul Gamble, Benjamin Lee



Machine learning to **detect** and **characterize** synthetic DNA sequences



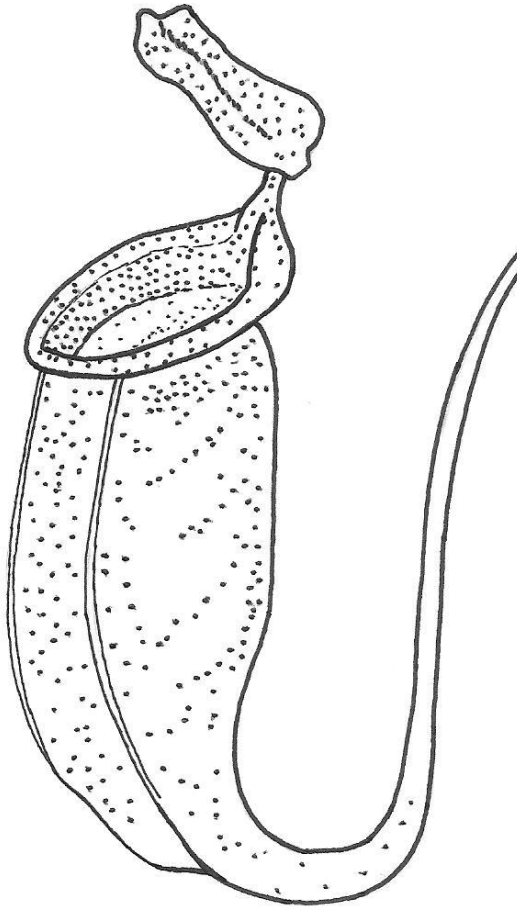
Prior Work

Current Research

Data Sources

Timeline & Deliverables

Future Aims & Collaboration



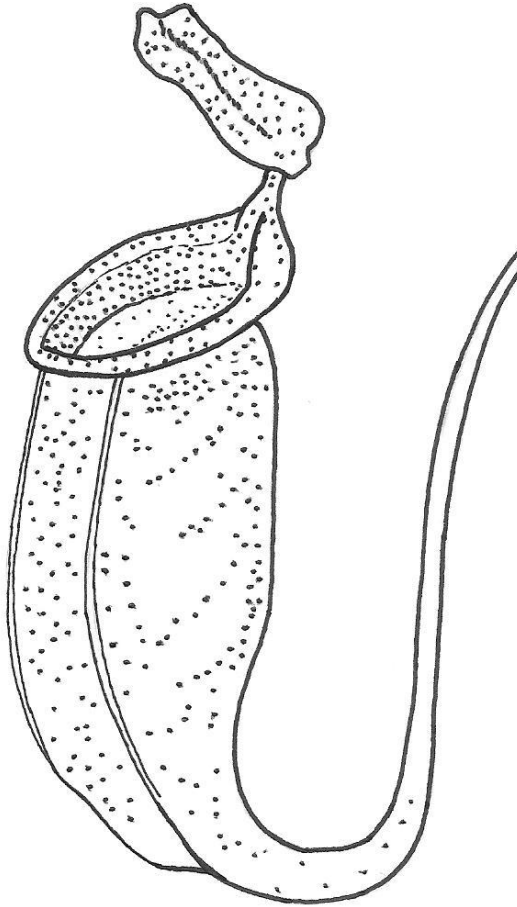
Prior Work

Current Research

Data Sources

Timeline & Deliverables

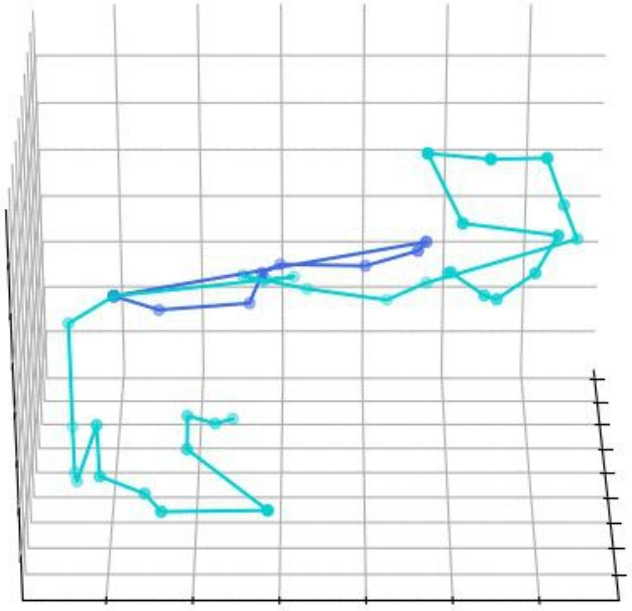
Future Aims & Collaboration



Prior Work: Cloning Boundary Detector

Cloning Boundary: Juxtaposition of sequences which occurs rarely in nature

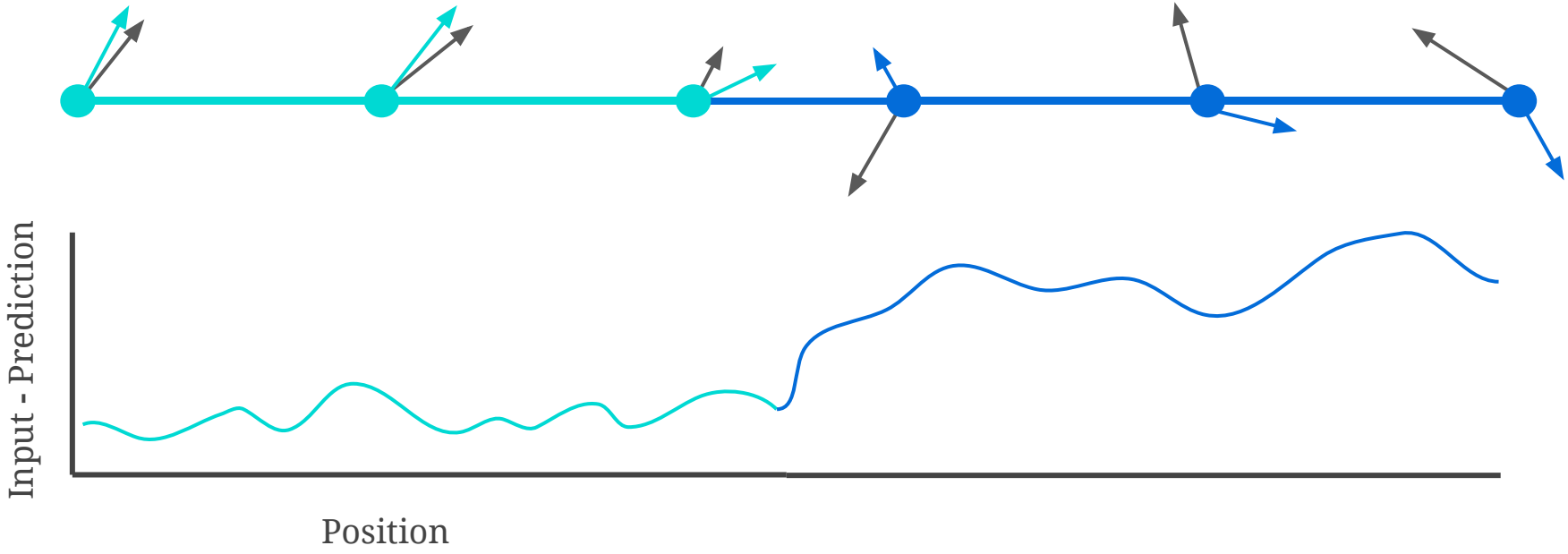
Prior Work: *k*-mer embedding



E. coli K-12 MG1655

mexY
(BAA34300.1)

Prior Work: Dissimilarity Metric



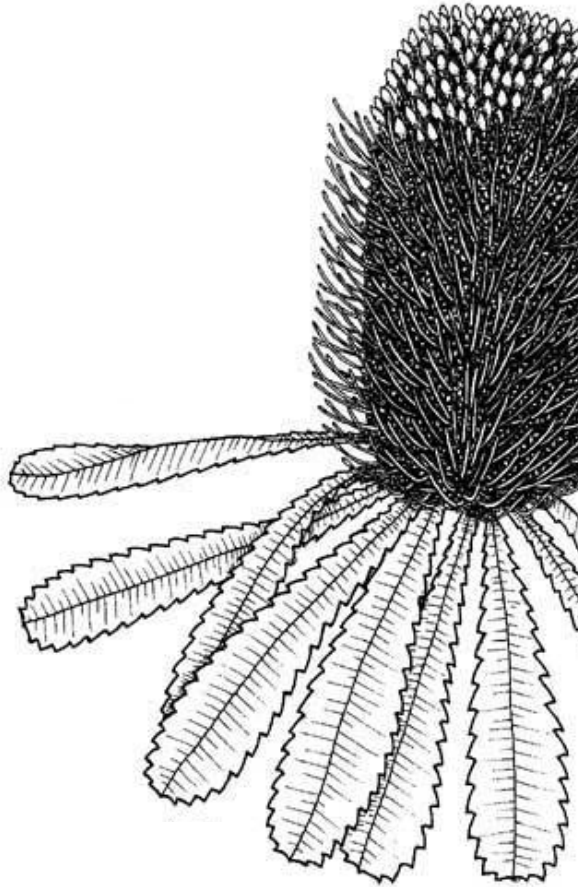
Prior Work

Current Research

Data Sources

Timeline & Deliverables

Future Aims & Collaboration



Focus Areas

1. Robustness to codon optimization
2. Alternative featurizations
3. Natural boundary recognition
4. Ensembling and meta-models
5. Boundary scale effects

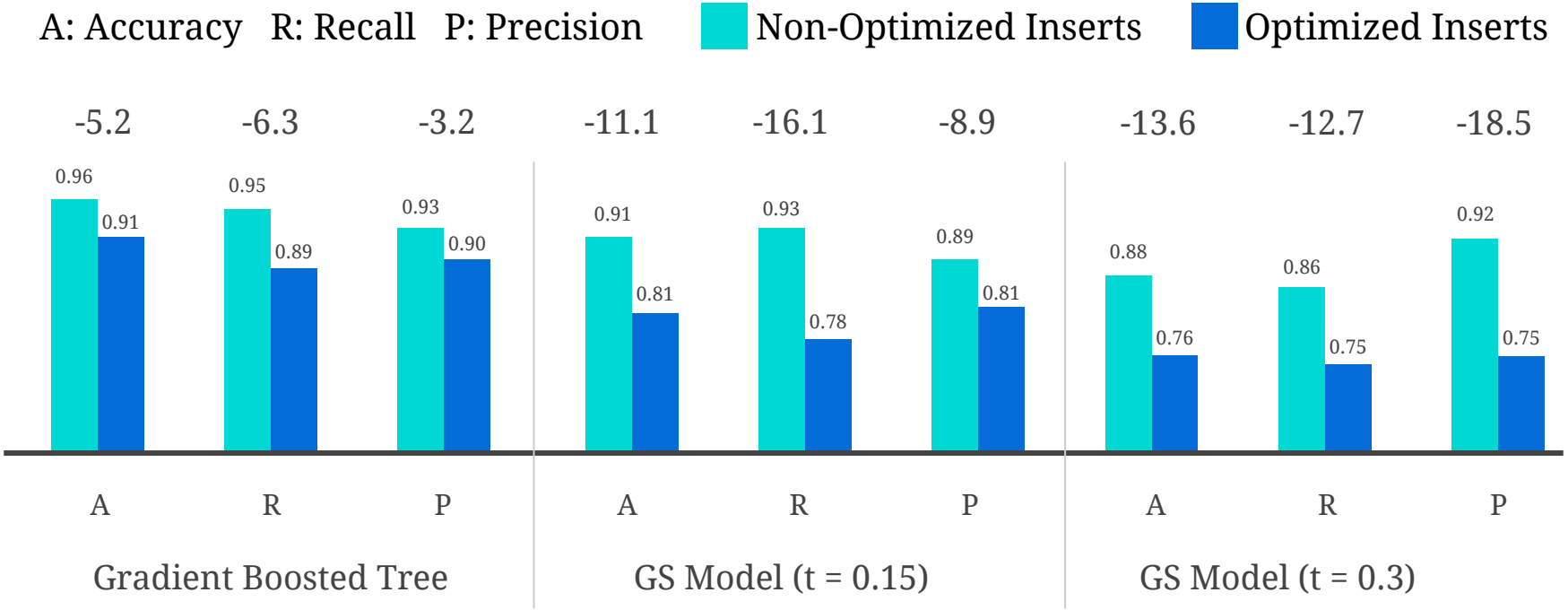
1. Codon Optimization

Codon usage bias:

- Powerful signal for sequence source determination
- Optimization distorts signal
- Can degrade classifier performance

1. Codon Optimization

A: Accuracy R: Recall P: Precision



1. Codon Optimization: Freqgen

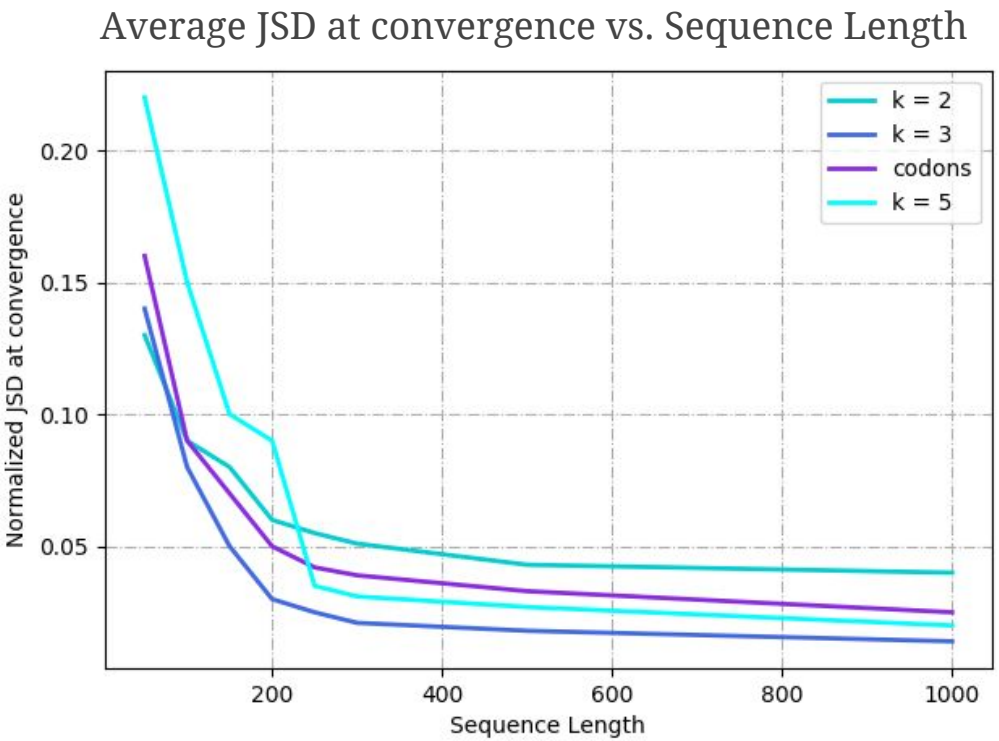
Freqgen: Genetic algorithm for generic *k*-mer frequency optimization using Jensen-Shannon Divergence

For discrete probability distributions P and Q:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$$

$$M = \frac{1}{2}(P + Q) \quad D_{\mathrm{KL}}(P \| Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)},$$

1. Codon Optimization: Freqgen



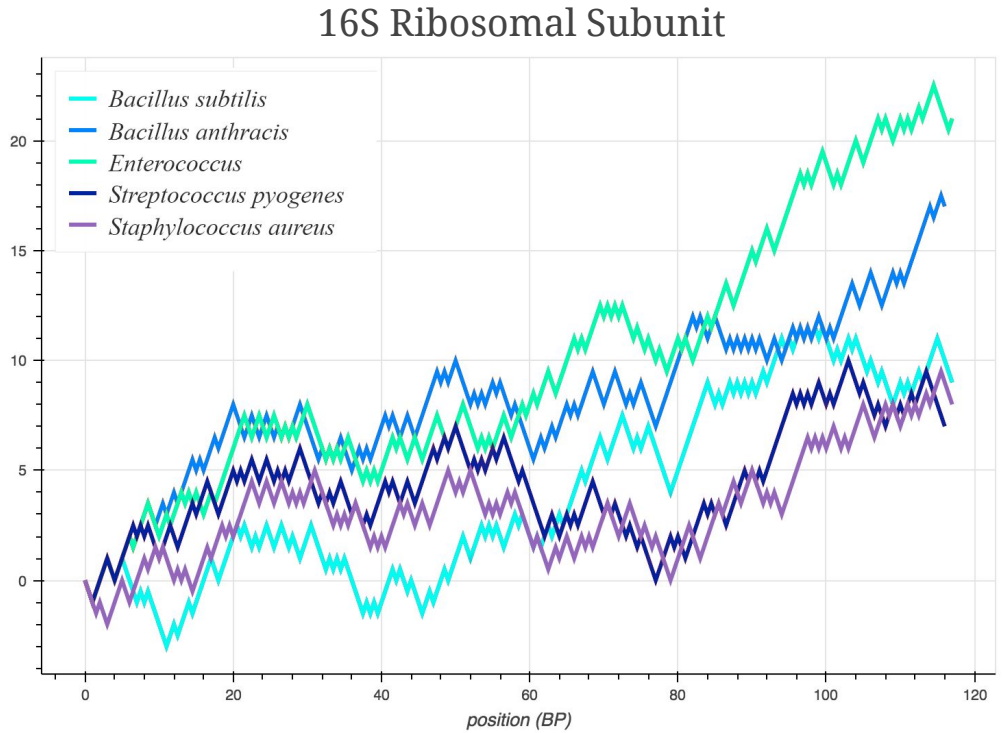
2. Alternative Featurization: BLAST

BLAST-based classifiers were found to be highly dependent on reference sequences

Initial experiments: High accuracy on test sequences in or near reference set, random guessing on outside sequences

Proposed experiments: Broaden the reference set, create several reference sub-sets, combine within a sorting model

2. Alternative Featurization: Sequence Plots



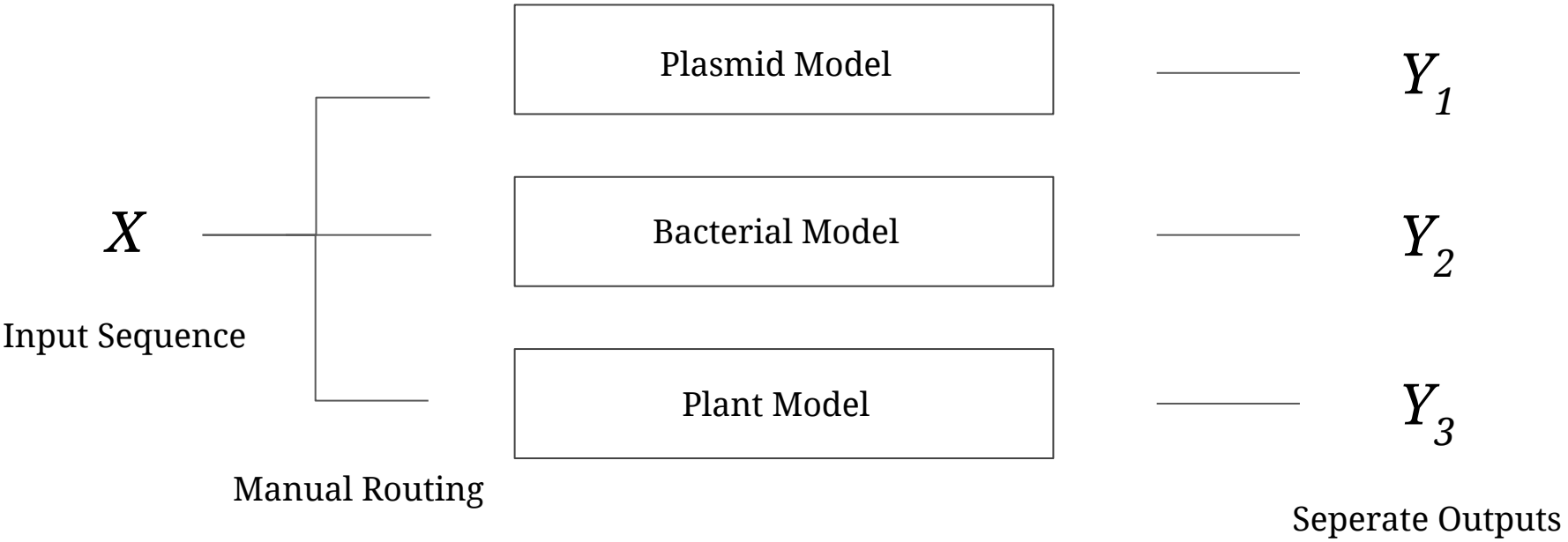
3. Natural Boundary Recognition

Juxtapositions that are rare in nature may still be naturally occuring: transposable elements, horizontal gene transfer

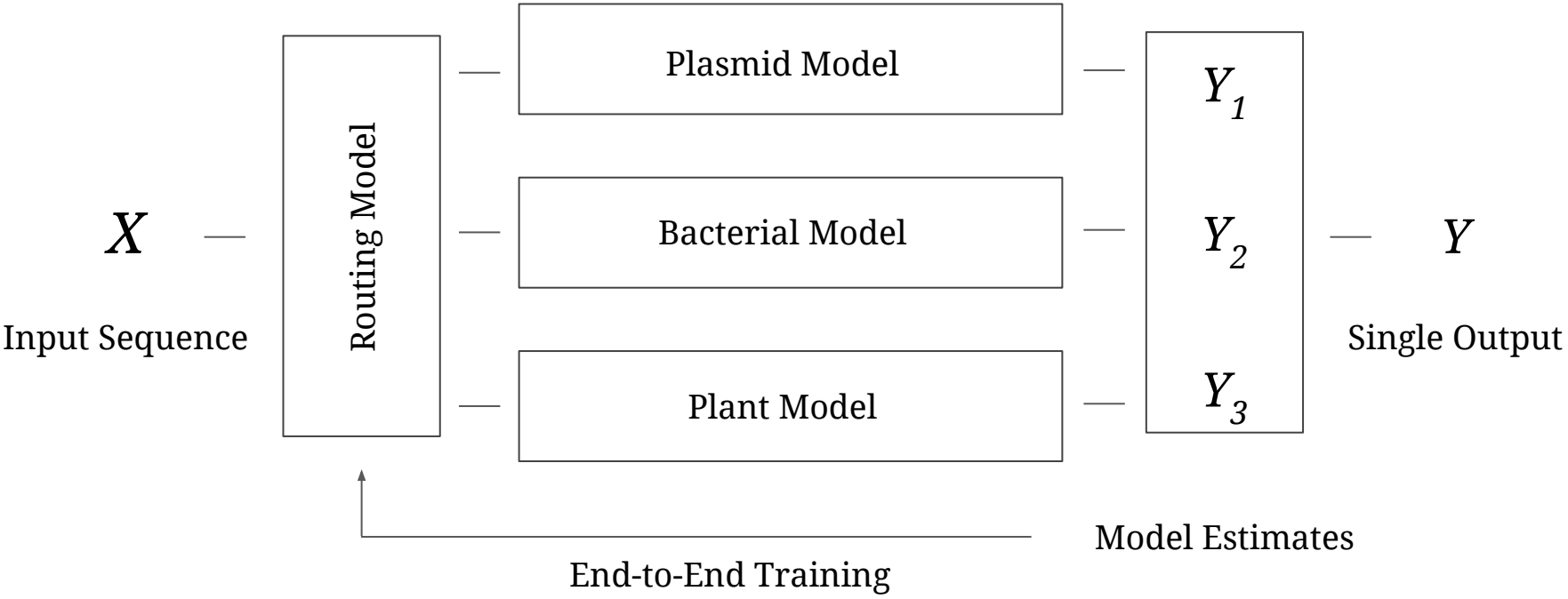
Initial experiments: GS Plant Model precision decreased from 84% to 36% on portions of the maize genome

Planned experiments: Expose model to natural boundaries during training, introduce a new class, layered models

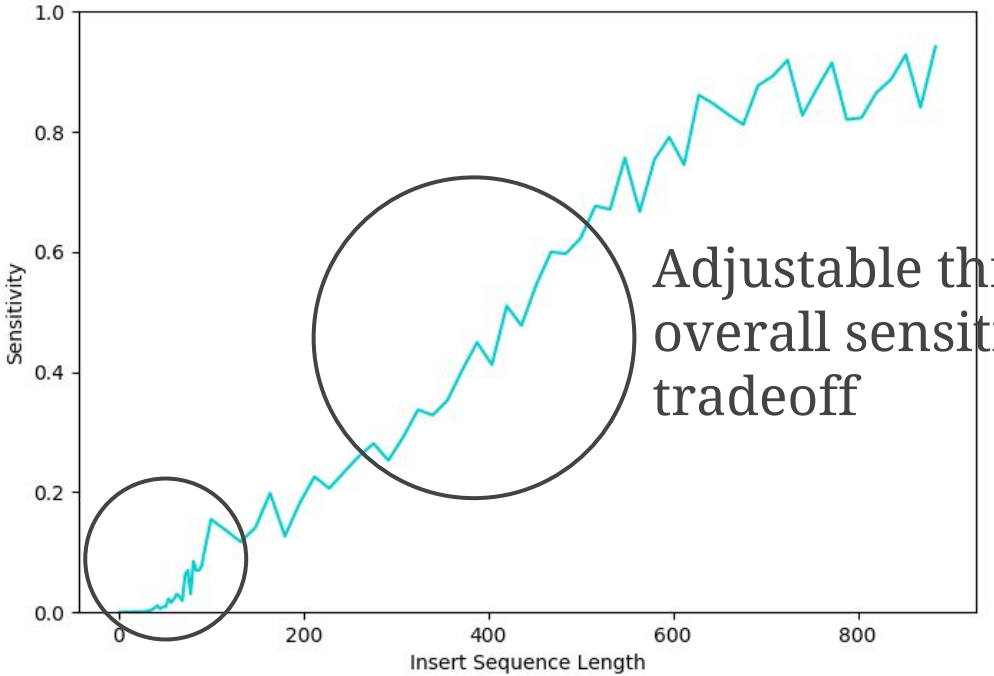
4. Ensembling and Meta-models



4. Ensembling and Meta-models



5. Boundary Scale Effects



Adjustable threshold: overall sensitivity/specificity tradeoff

Small insert non-linearity

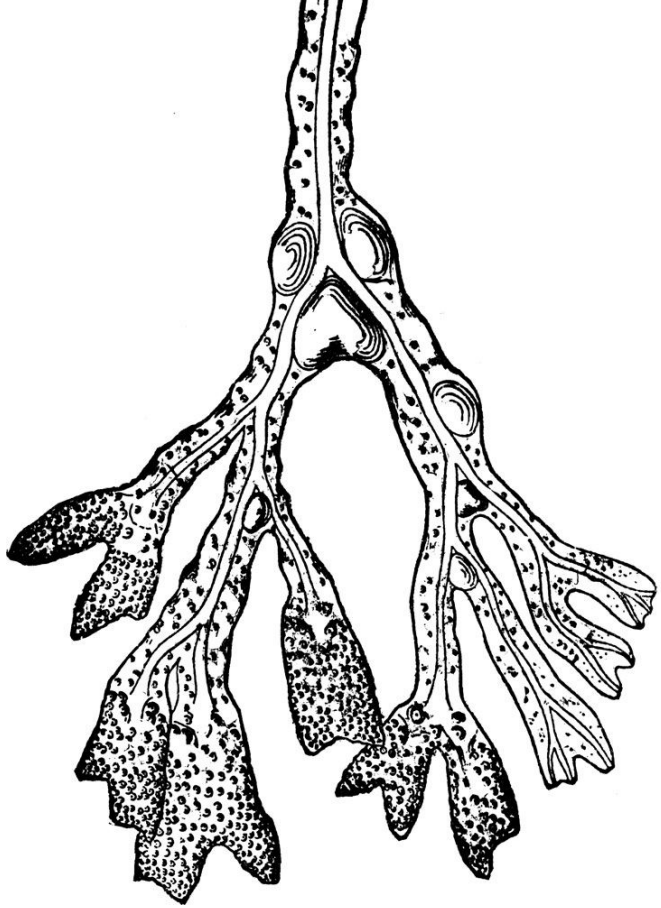
Prior Work

Current Research

Data Sources

Timeline & Deliverables

Future Aims & Collaboration



Current Data Sources

Virtual Synthetic Sequences

Backbones

RefSeq bacteria and plants
Commercial plasmids

Inserts

Antibiotic, metal, and herbicide resistance genes

Real World Data

Academic and Industry Collaborations

Collected Literature Sequences

Addgene plasmids

Data Augmentation Pipeline

Backbone Subsequence + Insert = Virtual Synthetic Sequence

Plasmids: Restriction enzyme cut sites, size limits

Bacteria and Plants: Random insertion, respecting known essential regions

Additional Datasets

Seeking **sharable collections** of modified sequences, particular emphasis on codon optimization methods

Dataset Creation: MNIST for synBio, a set of ‘modified model organisms’, various methods and inserts

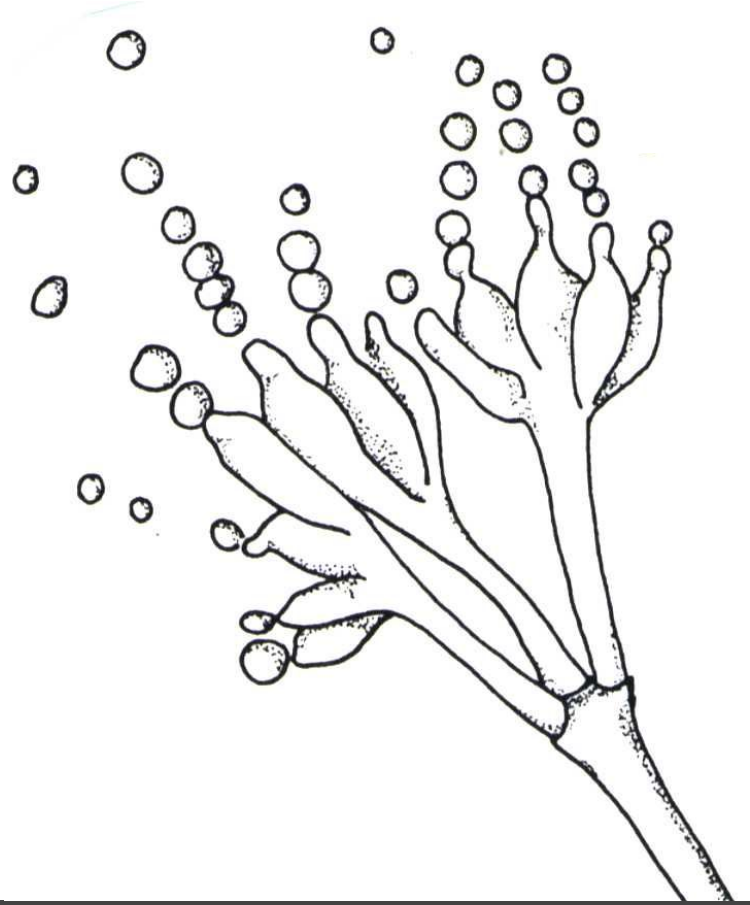
Prior Work

Current Research

Data Sources

Timeline & Deliverables

Future Aims & Collaboration



Deliverables

- Roundtable event early in the new year
- Open source code release (untrained models)
- Publications: Bioinformatics journals & ML conference
- Prototype: Synthetic DNA Geiger counter

NVIDIA Jetson + MinION Sequencer
= Portable Synthetic DNA Detection

Prior Work

Current Research

Data Sources

Timeline & Deliverables

Future Aims & Collaboration

Future Research Aims

Attribution: Engineering, sequencing and assembly techniques, lab and country of origin

Functional Characterization: Predict phenotypic effects from a purported modification


LAB41