

Deep Learning for Crop Gene Expression Prediction from Sequence Data

An Exploration of Chromosomal and Organellar Genomes

Sarah Davis
Greiner Lab Meeting
August 20th 2021

Canada NRC·CNRC amii

Presentation Overview

1 **Introduction**
Problem statement and motivation

2 **Data**
Exploration of datasets provided by the NRC

3 **Methods**
Data preprocessing and experimental set-ups

4 **Results**
Performance of each of the experimental set-ups

5 **Discussion**
What do the results tell us?

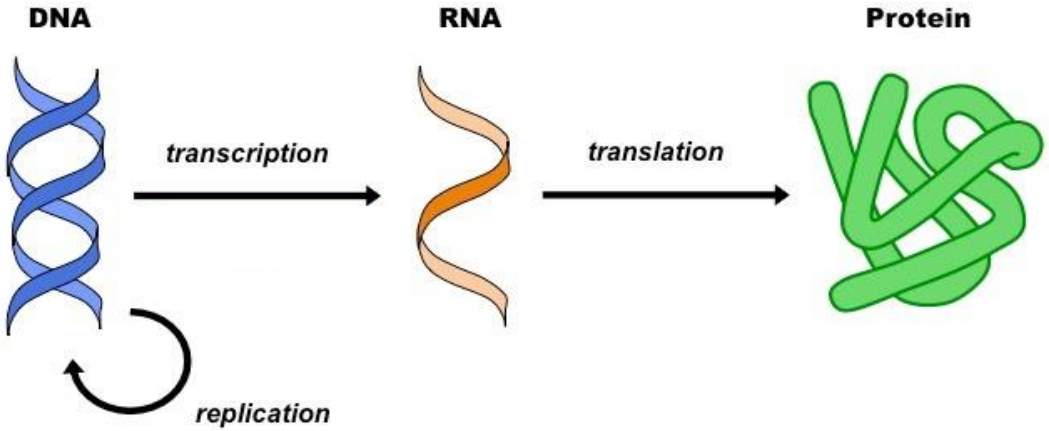
6 **Future Work**
Plans for the next two years



Introduction

Problem statement and motivation

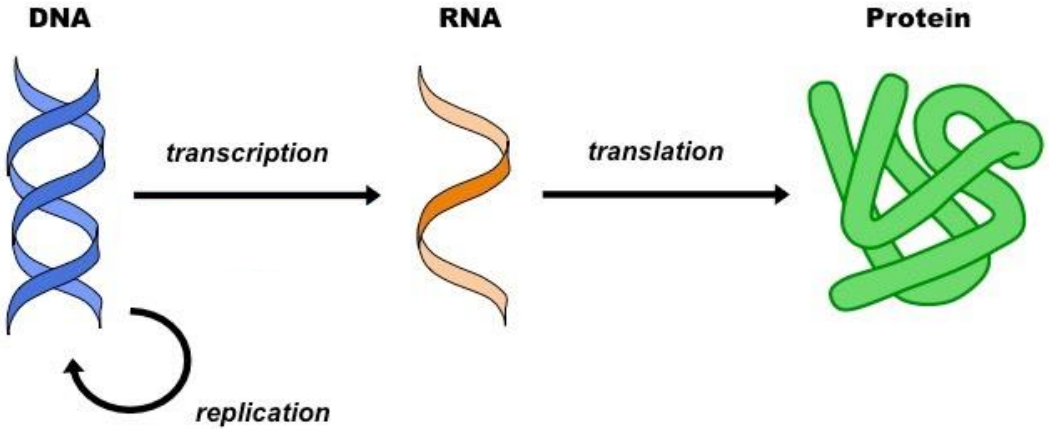
What problem are we trying to solve?



Increasing Usefulness

Increasing Measurement Difficulty

What problem are we trying to solve?

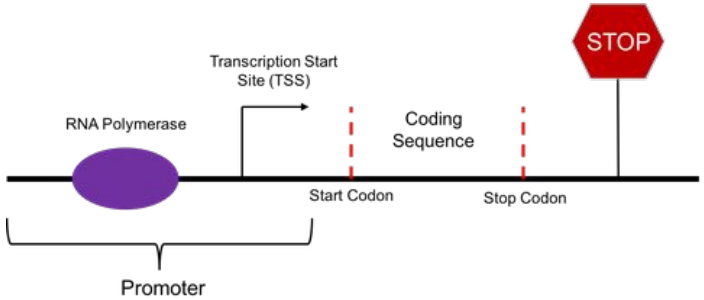


Use  to predict change in , which approximates  **

** in Canadian protein crops!

Plan

- Train a model to predict if a sequence acts as an enhancer
  - Enhancer: sequence of nucleotides that increases a gene's expression levels



- Use genomic data from different organisms and cellular sources

- Create a baseline model (which most previous literature lacks), use a number of different techniques for training, including interesting architectures and transfer learning

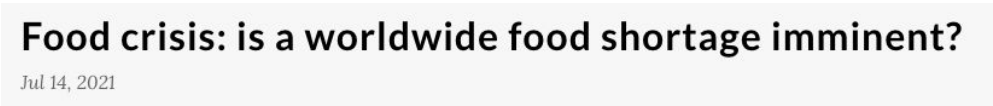
Potential Downstream Uses

- With global climate change and an increasing population, we need to be smart about how we make food











- Luckily, there are a number of interesting avenues to explore for tackling these issues!

Potential Downstream Uses



Potential *BNGRW* gene enhancers:

SQNC1: `...TAGAACCCTAGATCGCAT...`

SQNC2: `...TAGAACGGTAACACGCAT...`

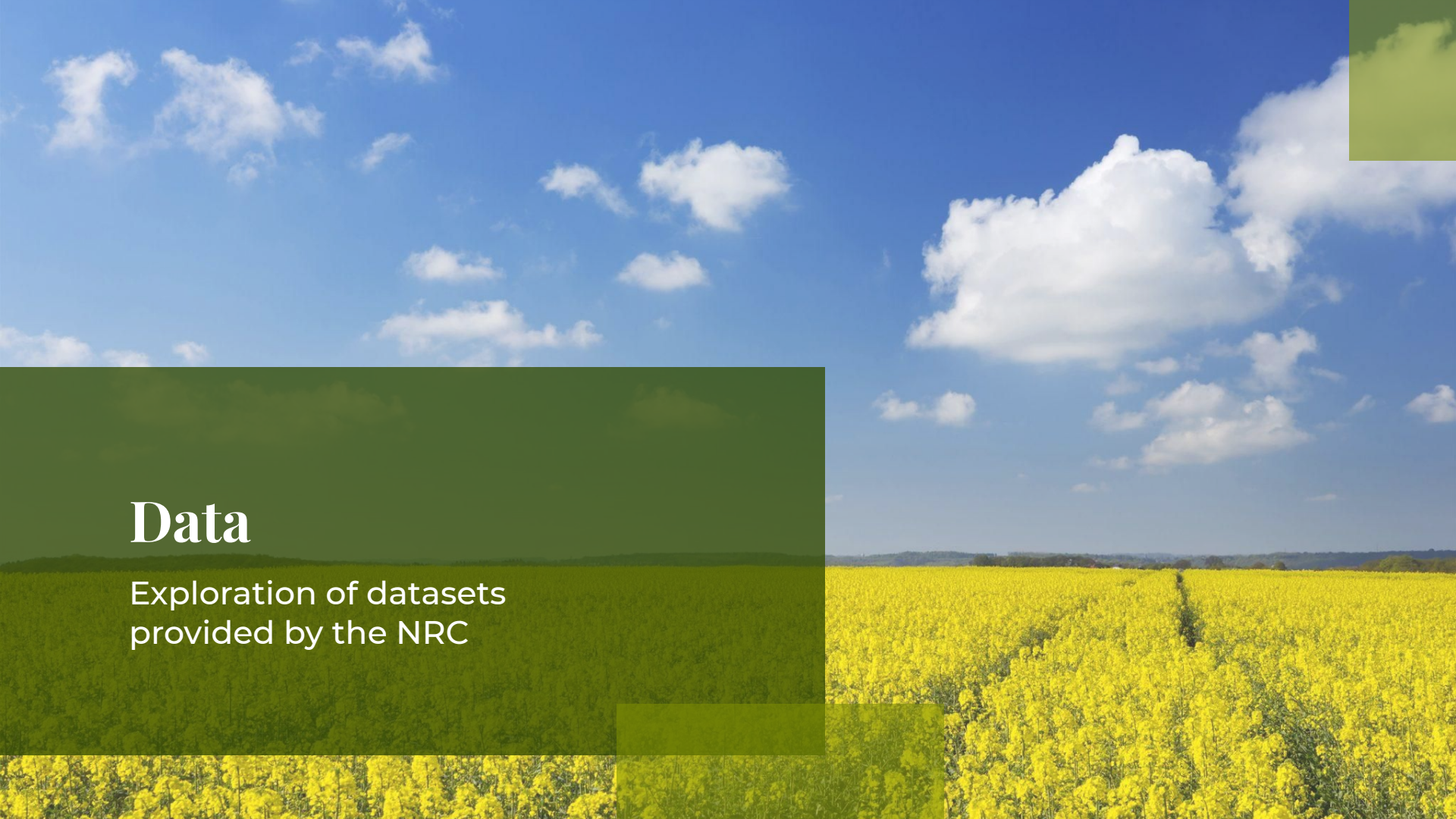
Model

SQNC1: `0.98 log2-FC`

SQNC2: `0.26 log2-FC`







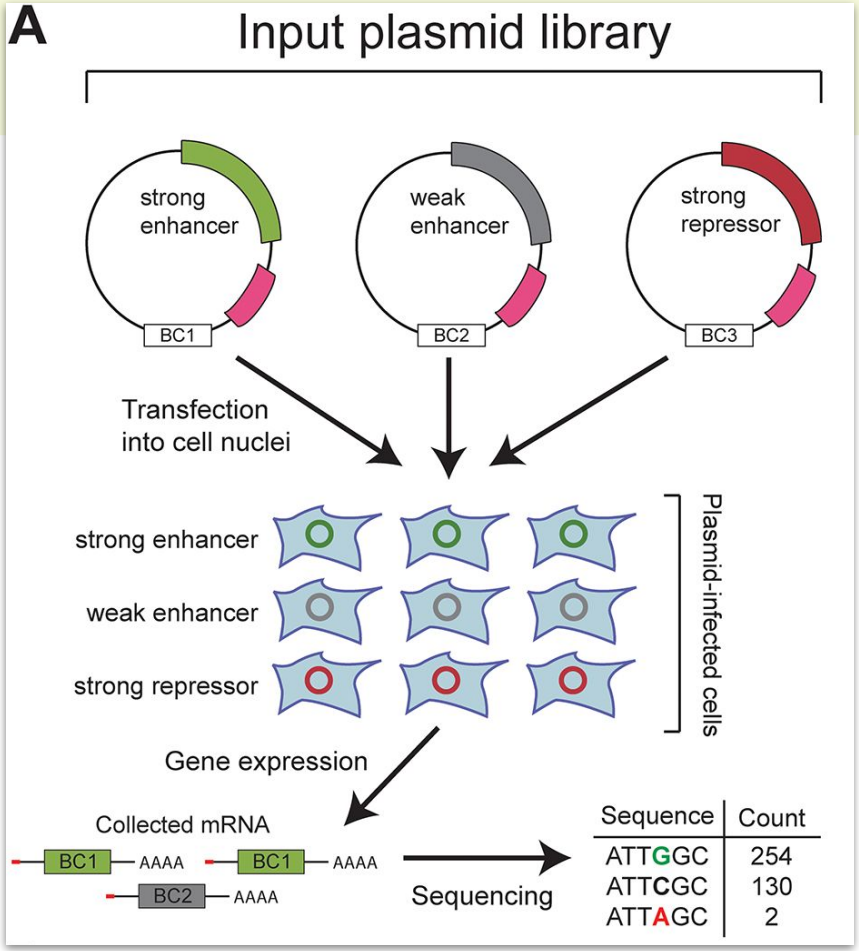
Data

Exploration of datasets
provided by the NRC

STARR-Seq

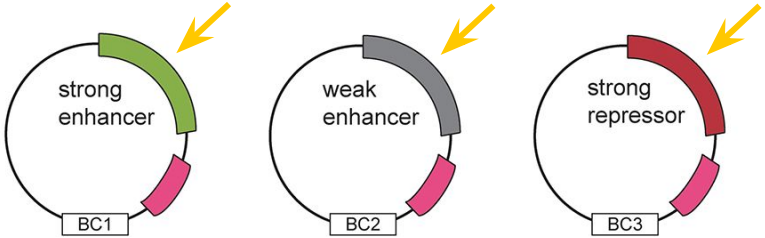
- Technique for generating the data we need
- Outputs pairs of:
  - 145-bp sequence
  - Associated expression change (target)

```
a = norm baseline expression
b = norm treatment expression
target = log2(a/b)
```



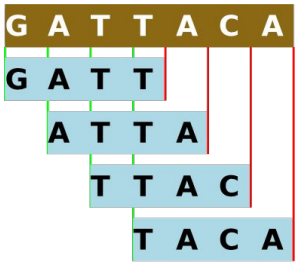
Potential Enhancers

- Where do these enhancer/ repressor sequences come from?



- If you have a list of potential candidates, that's great!
  - But, it's much more likely you're trying to discover candidates

- Solution: shear genome to create 145-bp sequences with 5 bp step size
  - Simple example:




```
total base pairs   = 7
k-mer length       = 4   (ours = 145)
step size          = 1   (ours = 5)
Result sequences   = 4
```

Two Datasets

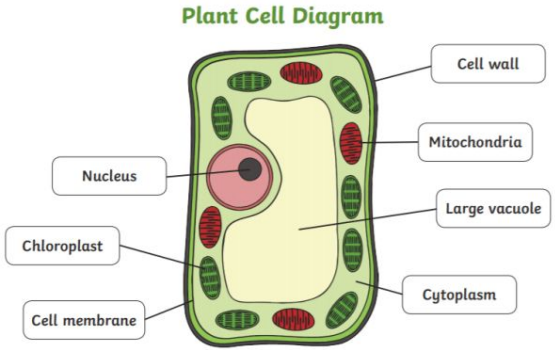
- This summer I worked with two datasets, each from a different organism

- Canola (Brassica napus)
  - Organellar DNA
    - Mitochondria
    - Chloroplast



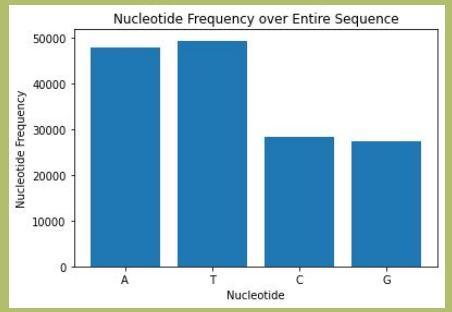
- Thale cress (Arabidopsis thaliana)
  - Organellar DNA
    - Mitochondria
    - Chloroplast
  - Chromosomal DNA
    - Chromosomes 1-5

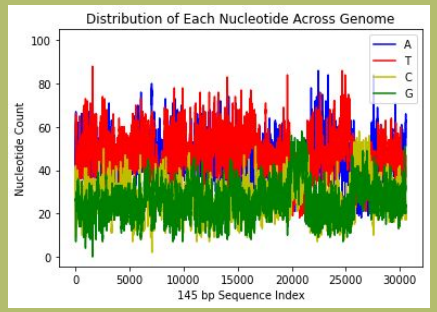


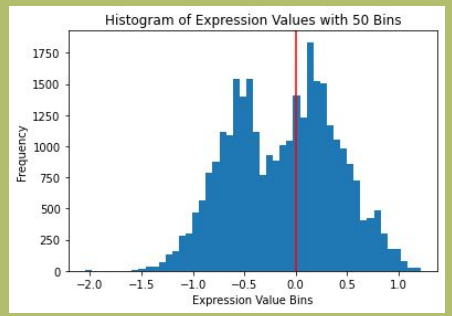


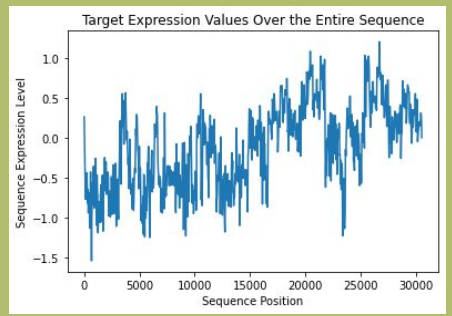
Canola Dataset

Chloroplast (nrows = 30544)

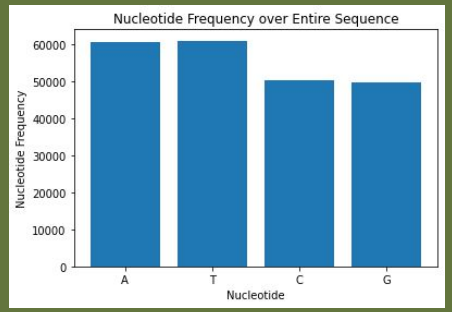


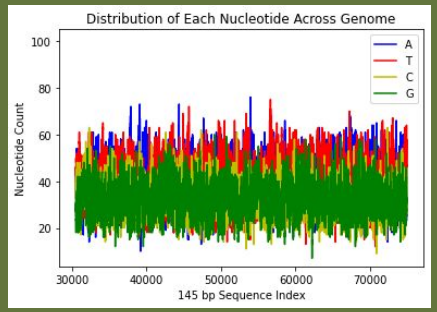


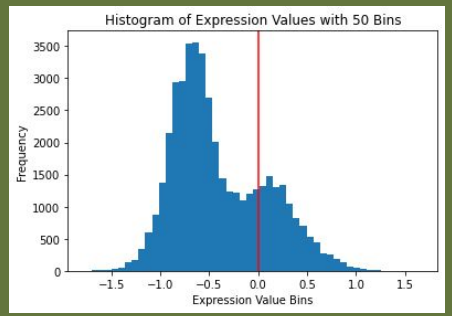


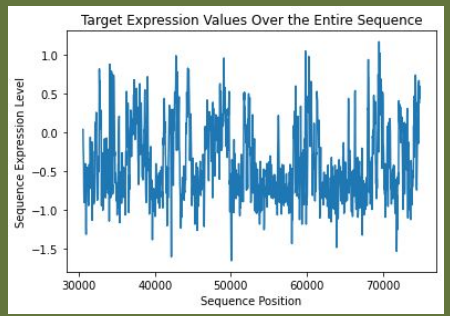


Mitochondria (nrows = 44342)



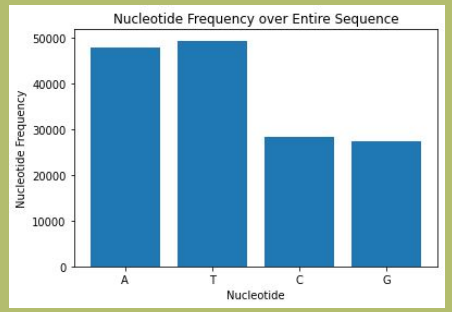


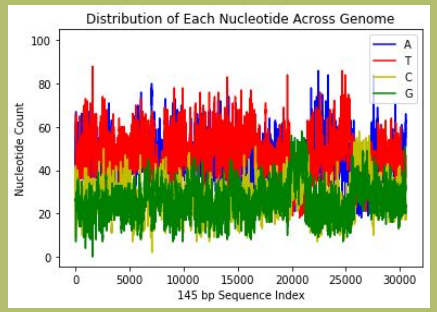


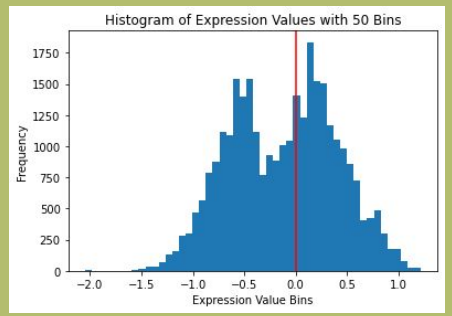


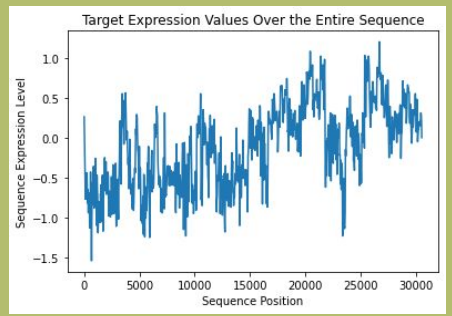
Canola Dataset

Chloroplast (nrows = 30544)

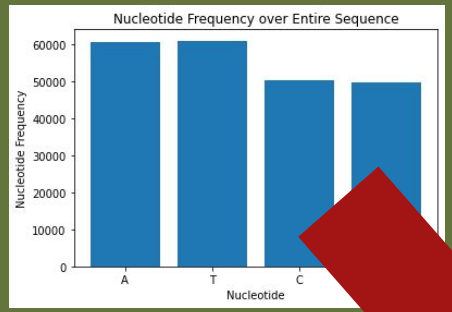


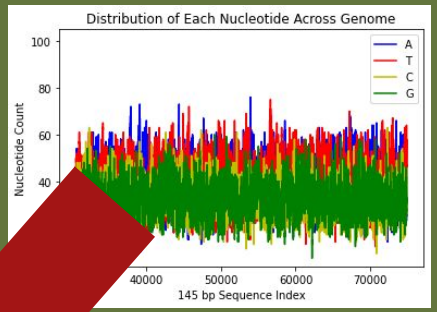


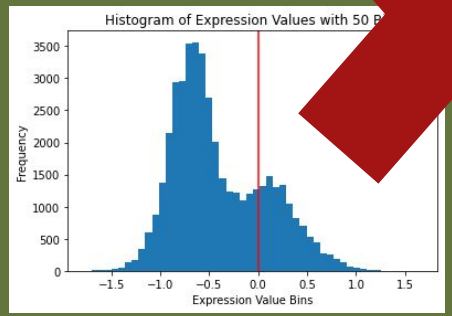


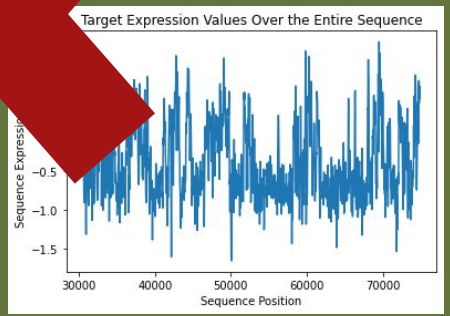


Mitochondria (nrows = 44342)



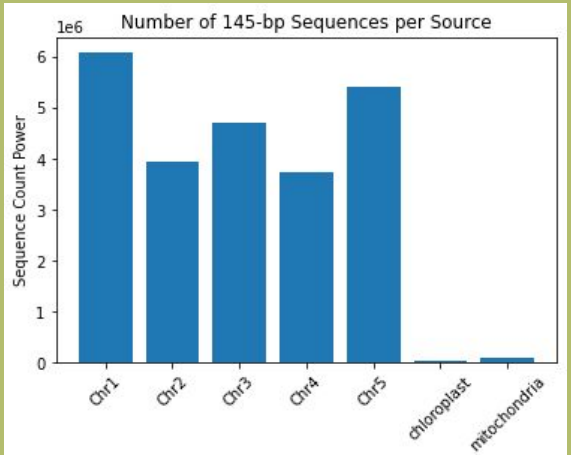






Arabidopsis Dataset

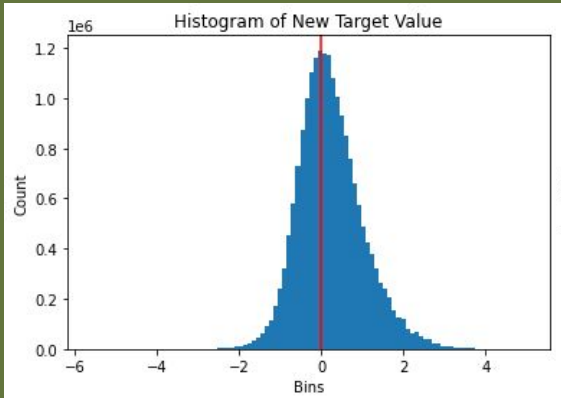
Sequence

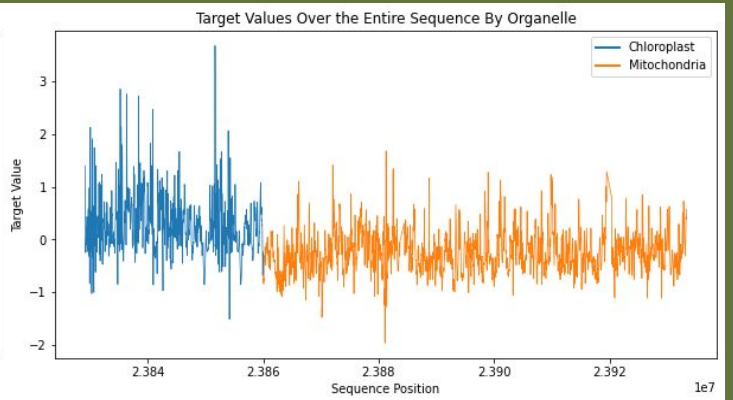


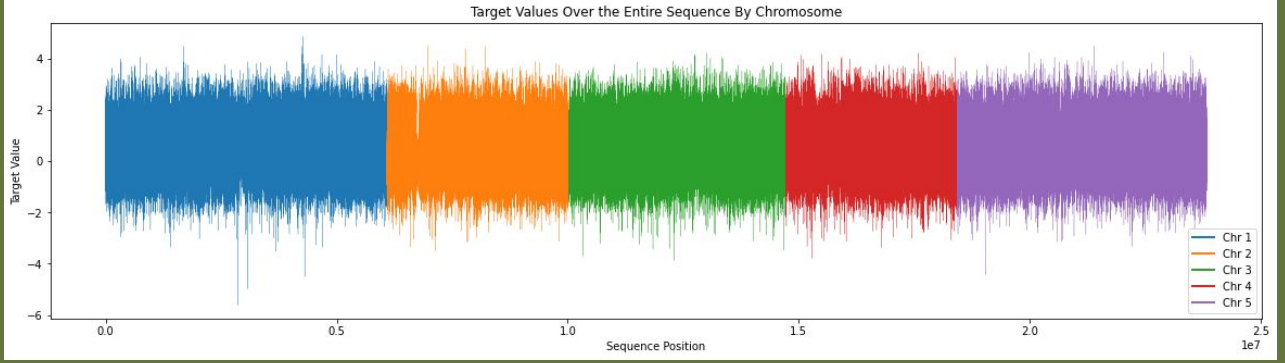
	A	C	D	G	K	M	N	R	S	T	W	Y
Chr1	9709674	5435374	0	5421150	53	76	163958	36	30	9697113	124	82
Chr2	6315640	3542973	0	3520763	10	5	2506	7	3	6316348	18	12
Chr3	7484757	4258333	0	4262704	0	2	5966	4	1	7448059	2	2
Chr4	5940546	3371349	1	3356090	0	1	3030	0	0	5914038	0	0
Chr5	8621974	4832252	0	4858759	0	0	10278	0	0	8652237	0	0
chloroplast	48545	28495	0	27570	0	0	0	0	0	49865	0	0
mitochondria	102463	82661	0	81608	0	0	0	0	0	100188	0	0

nrows = 23,933,351

Target Value









Methods

Data preprocessing and experimental set-ups

Frequency Baseline

- Similar to monograms, bigrams, and trigrams in NLP, in genetics we have mononucleotide, dinucleotide, and trinucleotide sequences
  - Short segments of 1, 2, and 3 base pairs

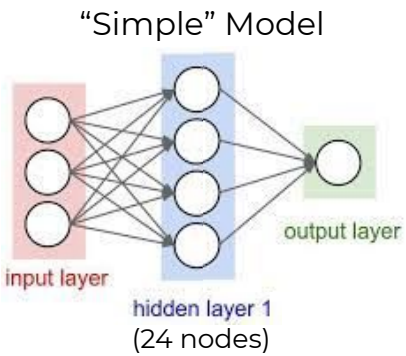
- The simple counts of these can act as features
  - MonoNT: 4 features
  - DiNT: 16 features
  - TriNT: 64 features

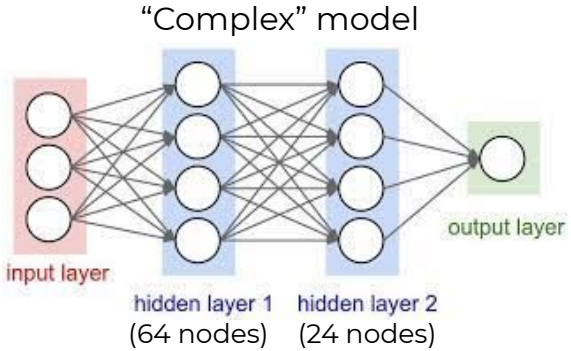
Target	Sequence
0.04	CGCCCCTACTCTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTT
0.08	CTACTCTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTT
0.06	CTACTATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATA
0.01	ATACATCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATATAATA(
-0.07	TCATATCTTGTTTGTATTGTCTAAAATAAACACGCAGCAATATTTTTTTTATCAAAAAAAAATTATAAATTATAATAGTATTTTTTTCTATTTTATTTTTATATATAATAGAAAA

Target	A	T	C	G
0.04	62	57	18	8
0.08	62	62	14	7
0.06	62	63	12	8
0.01	64	62	10	9
-0.07	65	59	11	10

Frequency Baseline

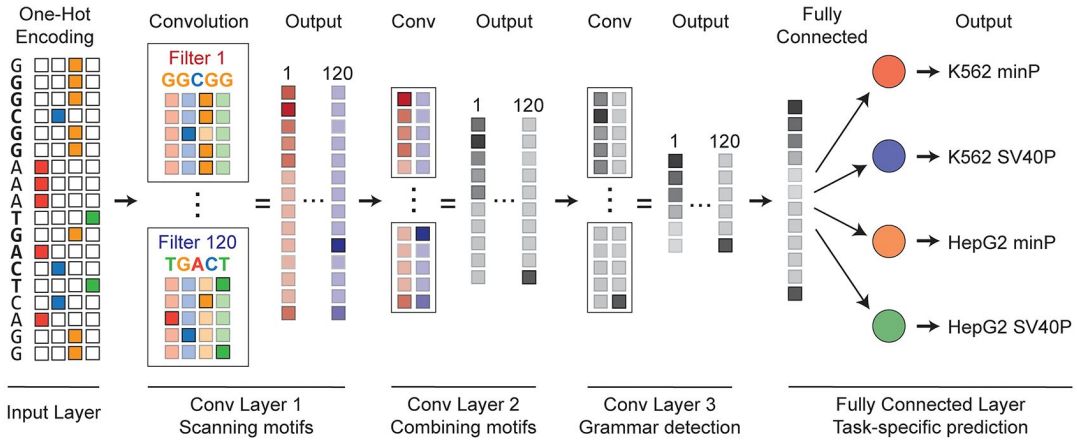
- Train two types of simple feed-forward neural networks to predict target column including every combination of NT sequences
  - MonoNT
  - DiNT
  - MonoNT + DiNT
  - TriNT
  - …





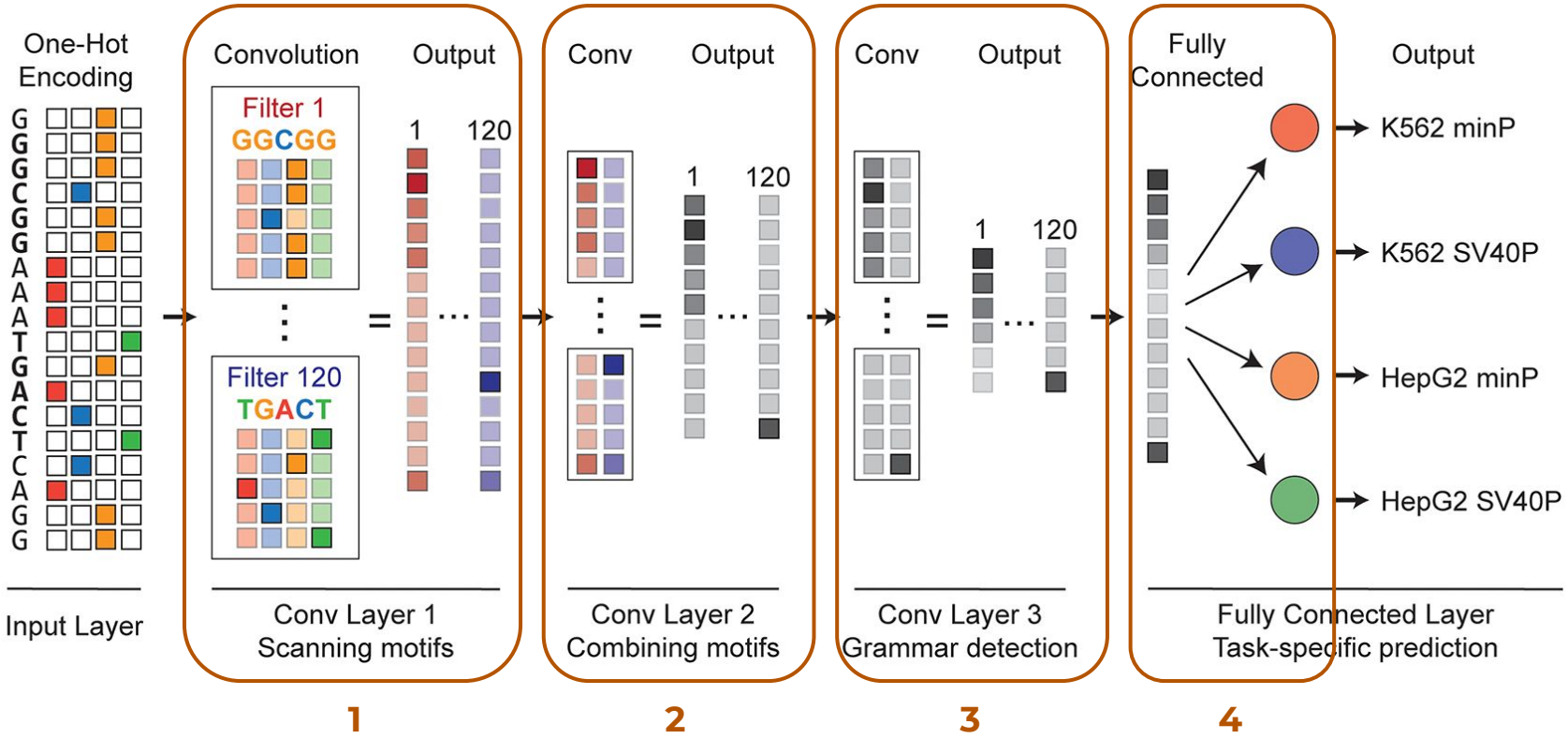
Literature: MPRA-DragoNN

- Uses convolutional models to detect motifs that are important in gene expression enhancement in human cell lines



- We want to use transfer learning to see if the low-level features learned in human cells are useful in a plant setting!

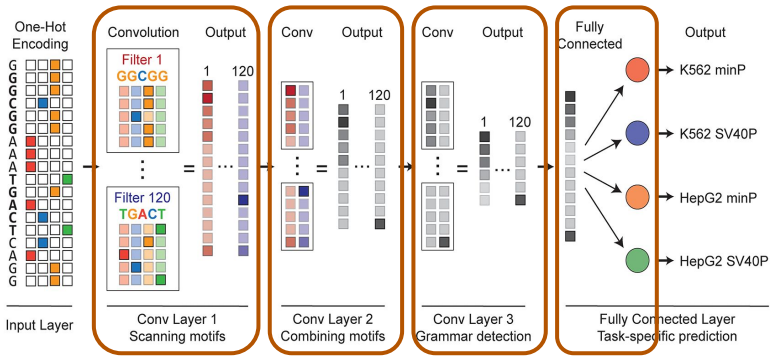
Experimental Set-Up



Experimental Set-Up

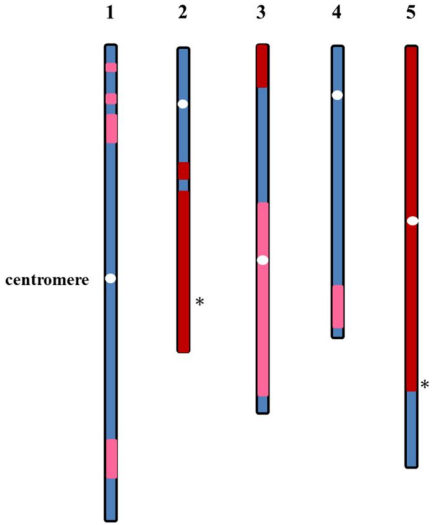
- Three possible treatments for each conv layer:
  a. Initialize random weights and train (0)
  b. Initialize MPRA-DragoNN weights and train (1)
  c. Initialize MPRA-DragoNN weights and freeze (2)

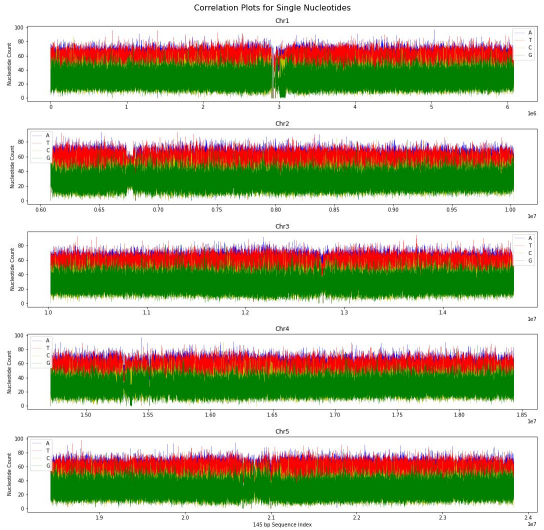
- We test the following setups:
  a. 000-
  b. 111-
  c. 222-
  d. 220-
  e. 221-
  f. 2222



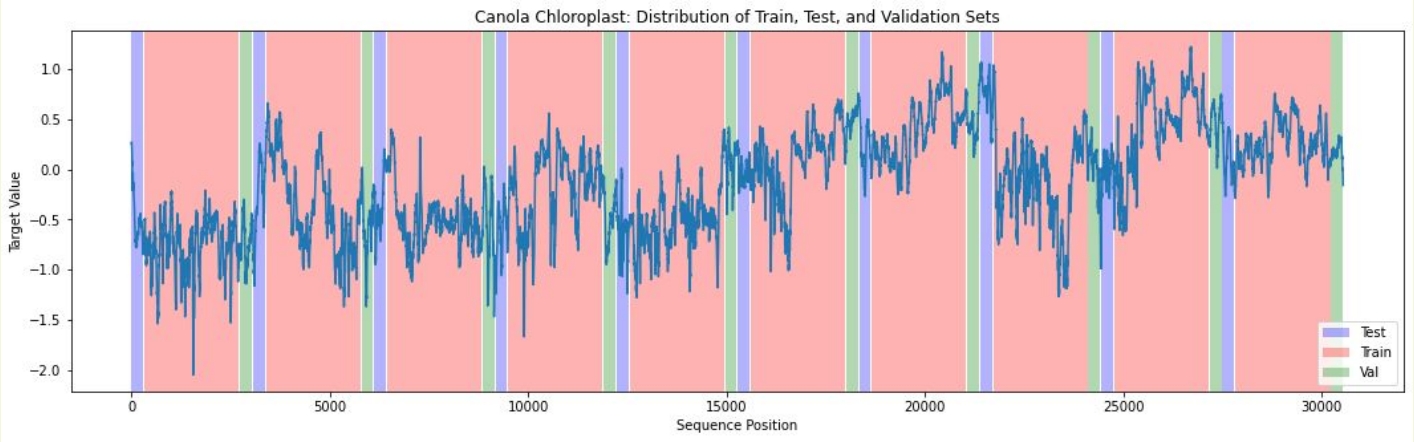
Train/Test/Val Splits

- Recall that our sequences contain 140bp overlap with each other
- Problem: If 97% of the sequence is overlapping with its neighbour...
  - Is it fair to have the sequence and its neighbour in different splits?
- Easy when you have chromosomes!
  - i.e. arabidopsis
- Not as easy when only working with the organellar genome
  - i.e. canola

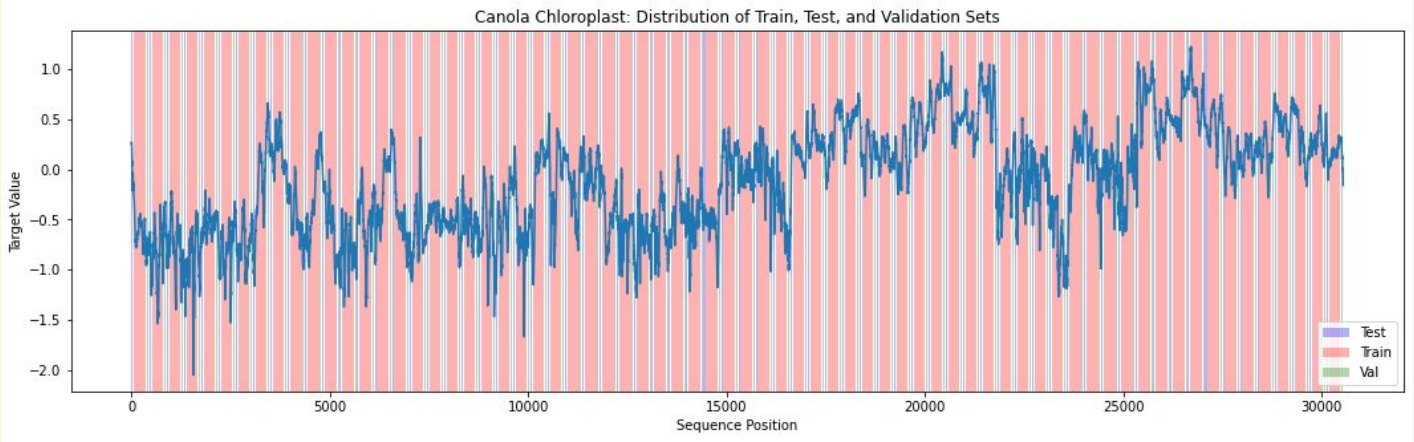


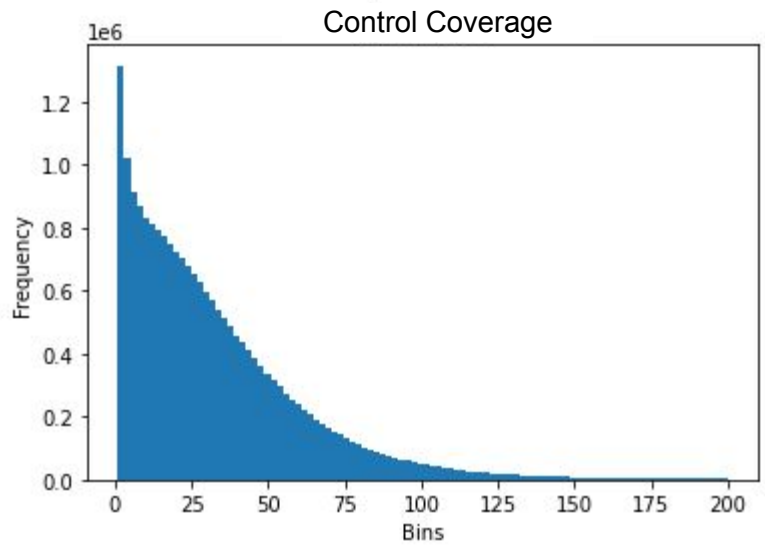


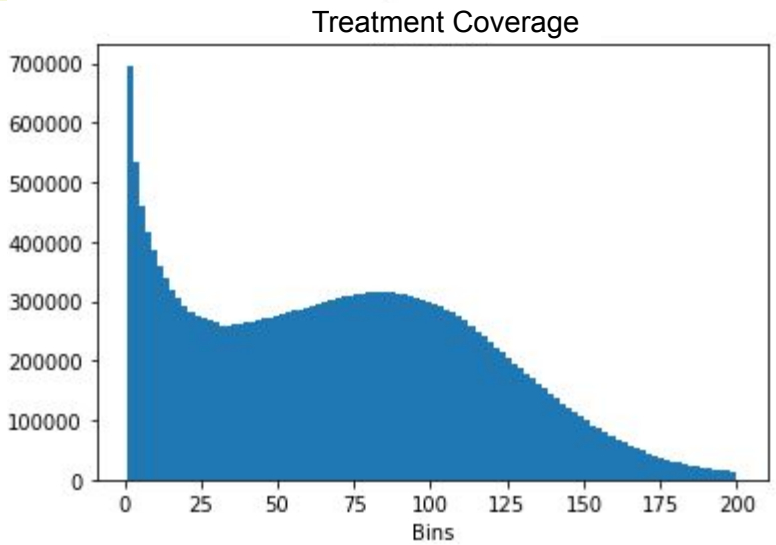
n_chunks = 10



n_chunks = 70




Control Coverage
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
25
50
75
100
125
150
175
200
Bins


Treatment Coverage
700000
600000
500000
400000
300000
200000
100000
0
0
25
50
75
100
125
150
175
200
Bins

Training

Settings:

Learning_rate = 0.01*
Batch_size = 512*

Patience = 20 (monitoring val_loss)
Max_epochs = 100

```
Epoch 19/100
1399/1399 [==============================] - 12s 8ms/step - loss: 0.4785 - val_loss: 0.4757
Epoch 20/100
1399/1399 [==============================] - 12s 8ms/step - loss: 0.4785 - val_loss: 0.4757
Epoch 21/100
1399/1399 [==============================] - 12s 8ms/step - loss: 0.4785 - val_loss: 0.4757
Epoch 00021: early stopping
```

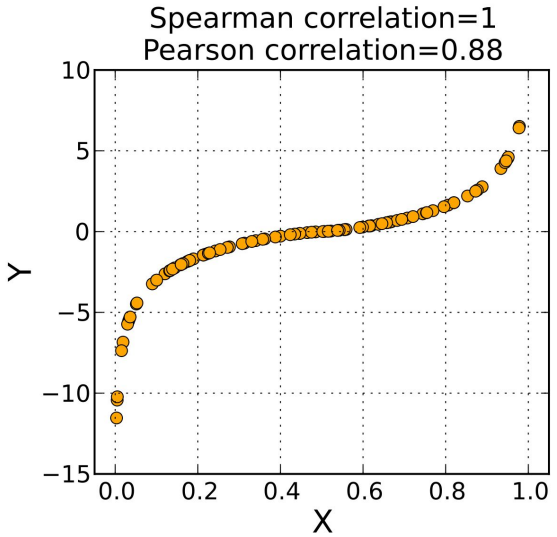
*as per MPRA-DragoNN, everything else was left as Keras defaults

How do we measure performance?

- We are interested in both a measure of magnitude of deviation between the prediction and reality, as well as a measure that considers ranking

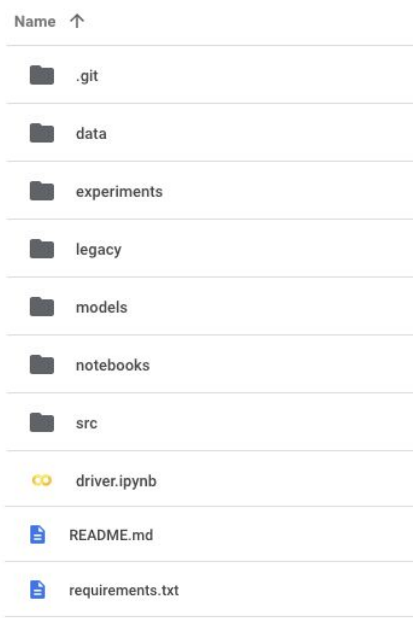
- We chose two options:
  - Pearson correlation coefficient (r^2)
    - Deviation magnitude
  - Spearman correlation coefficient (ρ)
    - Rank

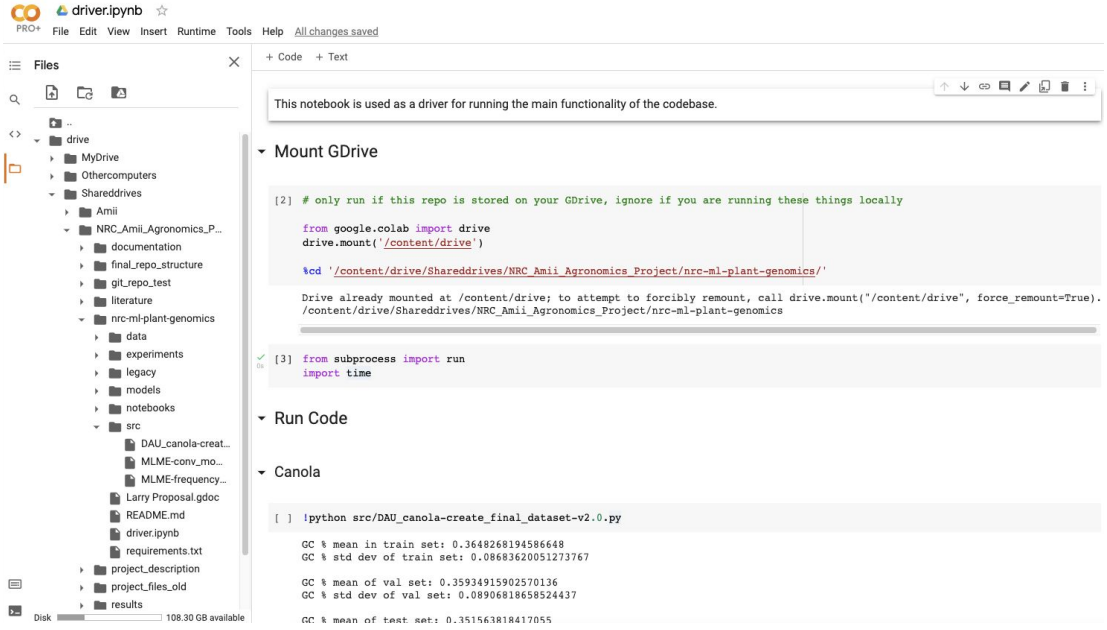
- Spearman coefficient tends to be slightly more lenient



Computing Environment

- Everything here was done entirely hosted on Google Colaboratory and Google Drive (plus Git/GitHub) with a Colab Pro+ account



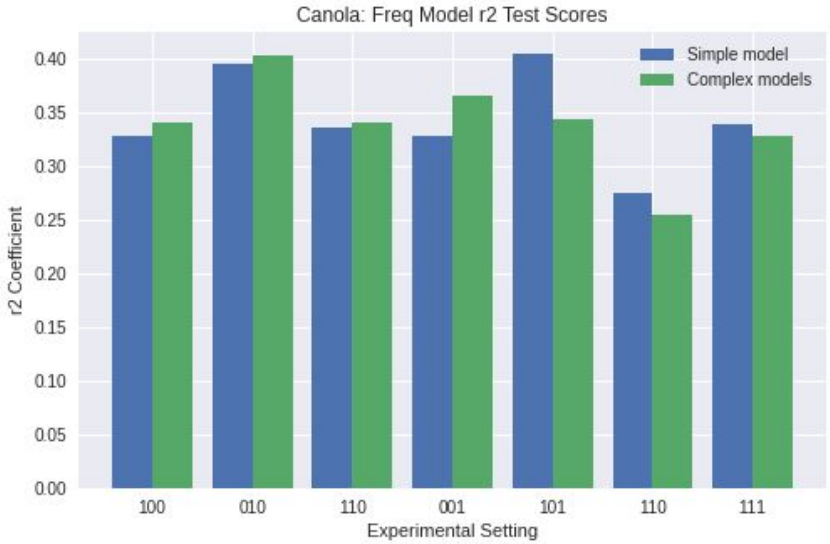


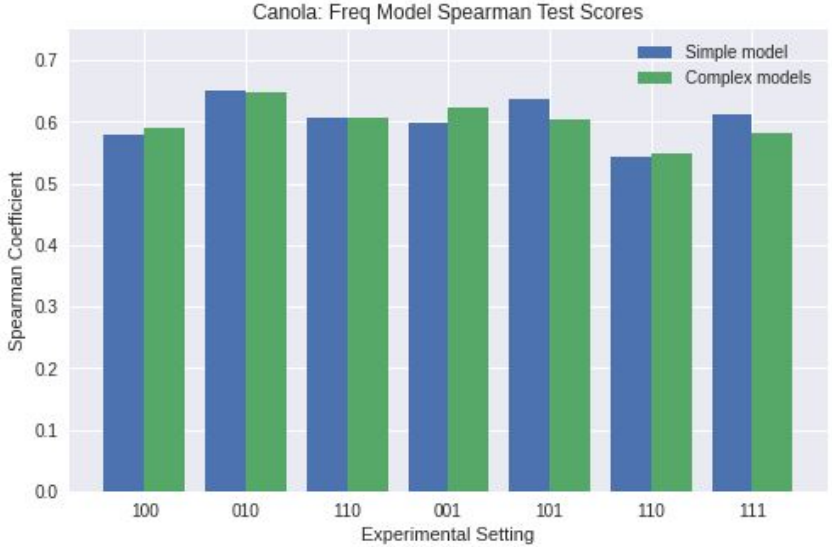


Results

Performance of each of the experimental set-ups

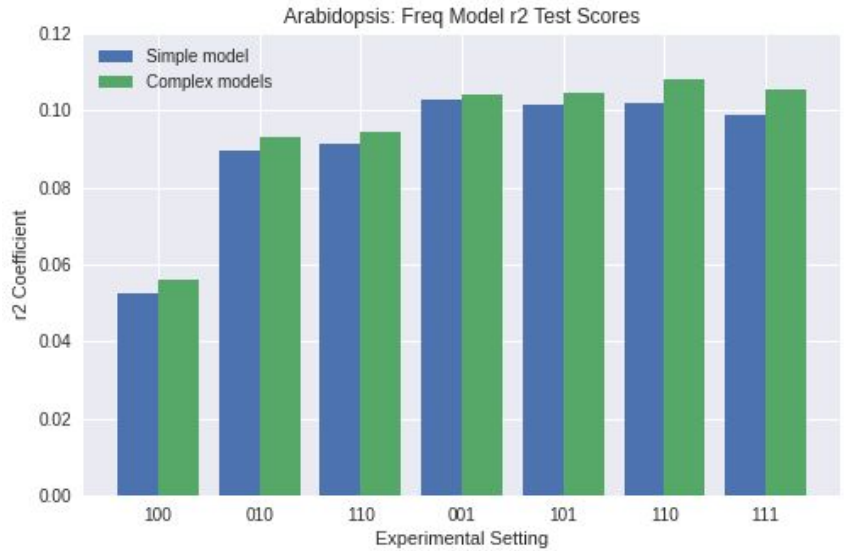
Canola Frequency Models

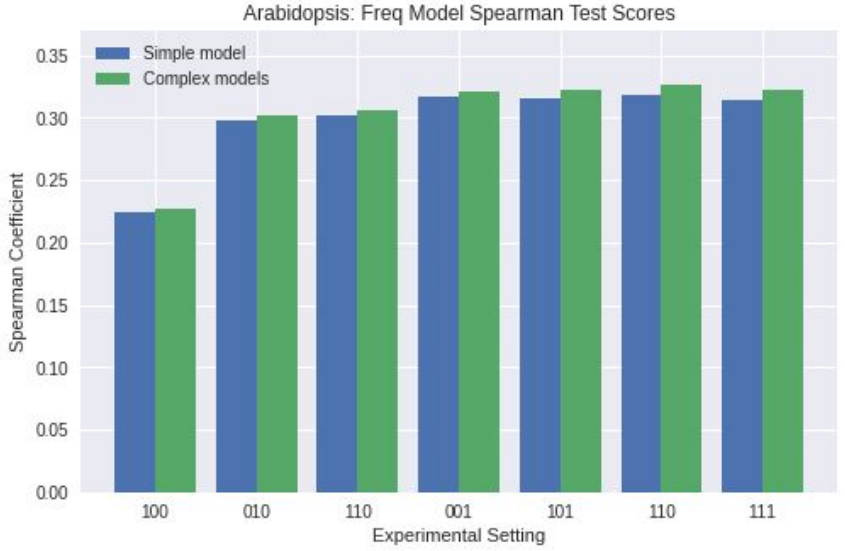




100 = MonoNT features
010 = DiNT features
001 = TriNT features

Arabidopsis Freq Model

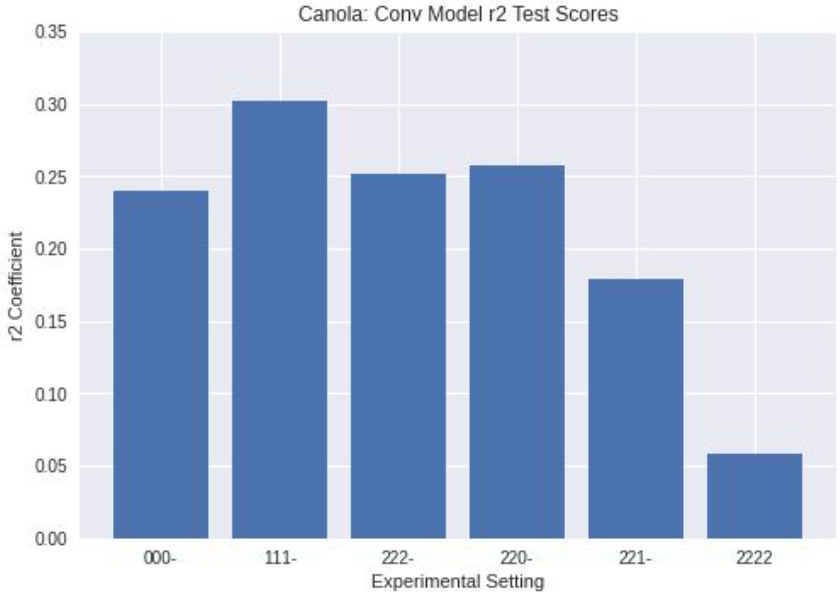


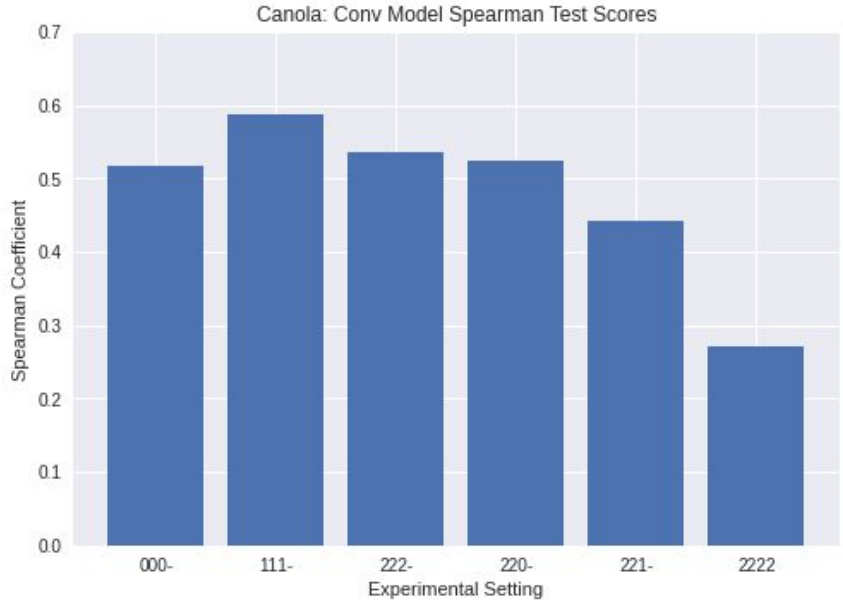


100 = MonoNT features
010 = DiNT features
001 = TriNT features

Canola Conv Model*

```
0 for init random weights & train
1 for load pre-trained weights and train
2 for load pre-trained weights and freeze
```

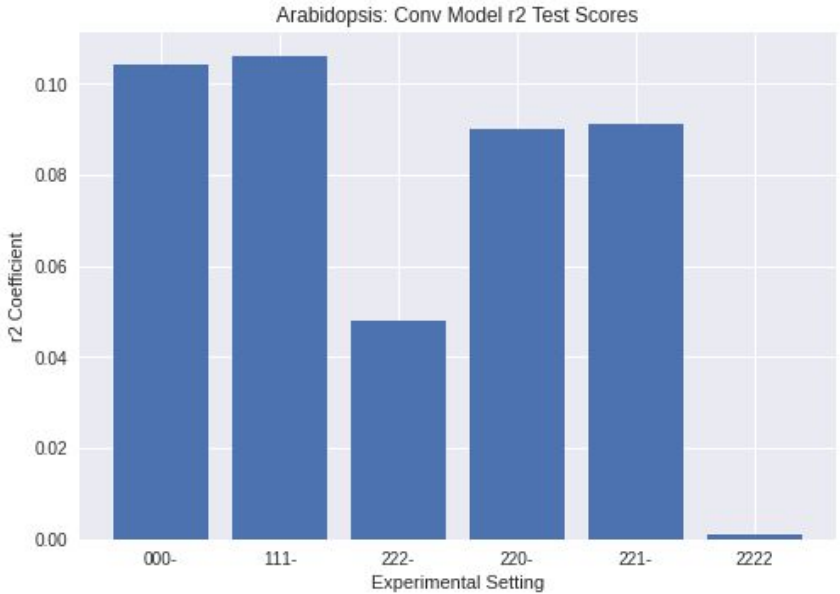


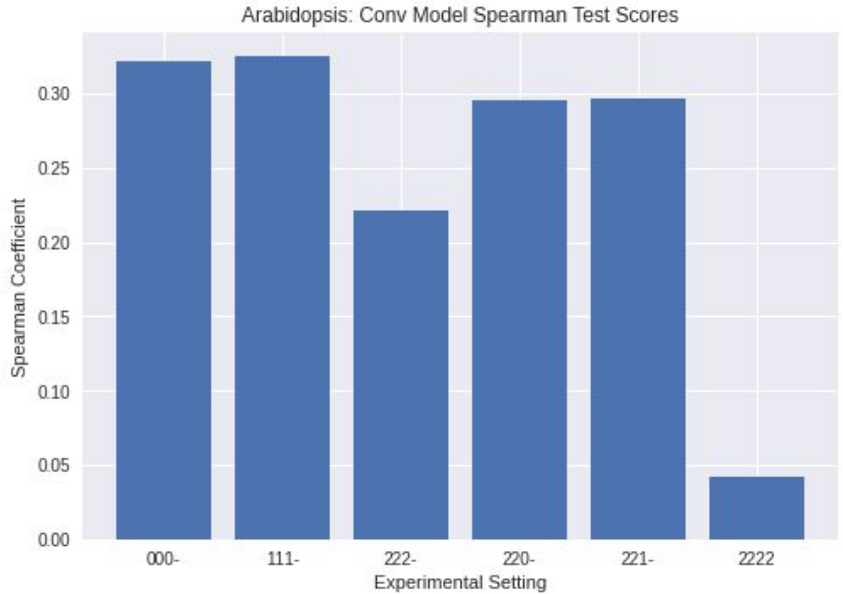


*Many, many more results not pictured from grid-searches and parameter tuning – used MPRA-DragoNN params for comparison

Arabidopsis Conv Model

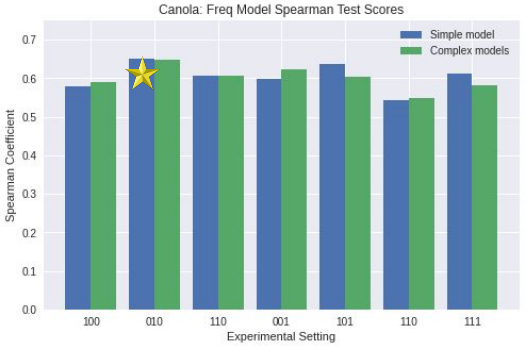
```
0 for init random weights & train
1 for load pre-trained weights and train
2 for load pre-trained weights and freeze
```



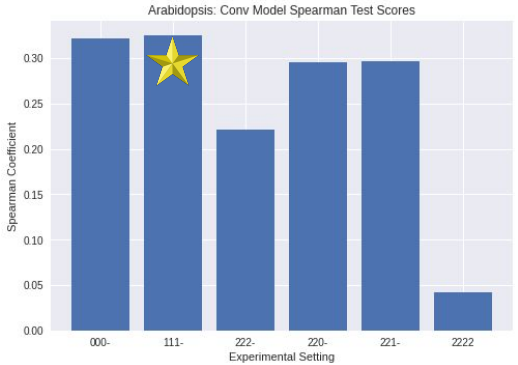


Best Scores*

- Organellar data:
  - r2: 0.404 (freq, simple, `101`)
  - Spearman: 0.652 (freq, simple, `010`)



- Chromosomal data:
  - r2: 0.108 (freq, complex, `110`)
  - Spearman: 0.325 (conv, `111-`)



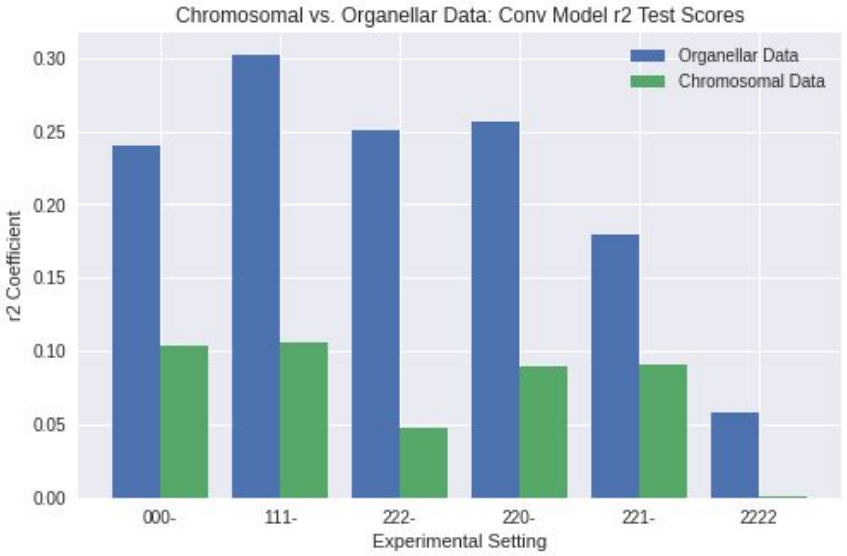
*No claims about statistical significance

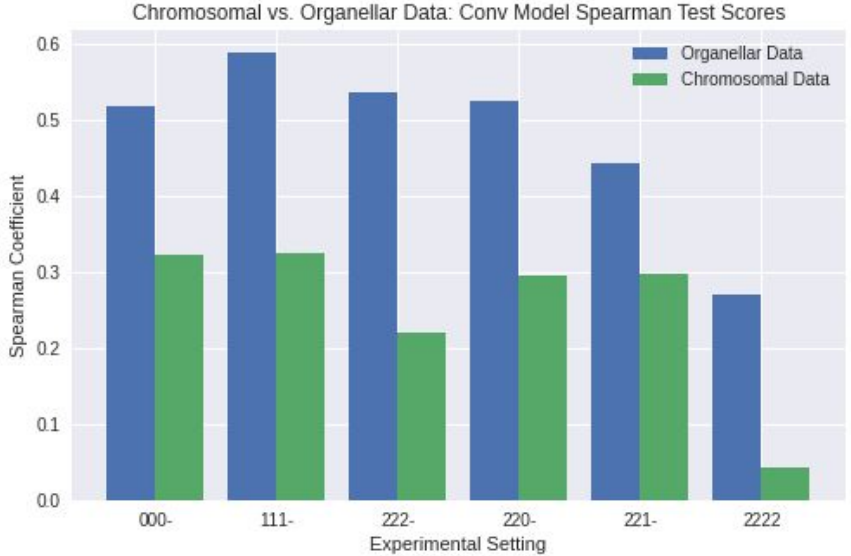


Discussion

What do the results tell us?

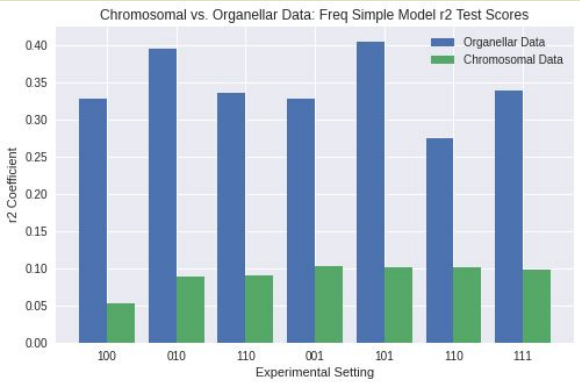
Organelle vs. Chromosomal Data (Conv)

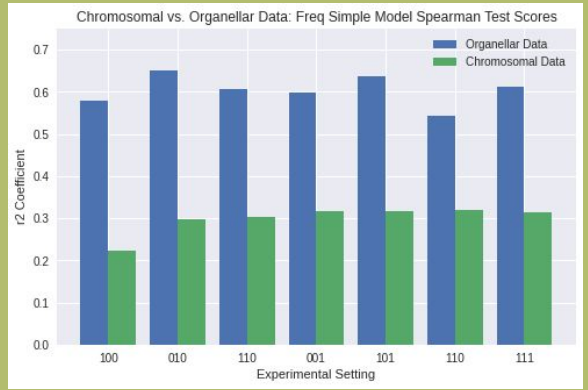




Organelle vs. Chromosomal Data (Freq)

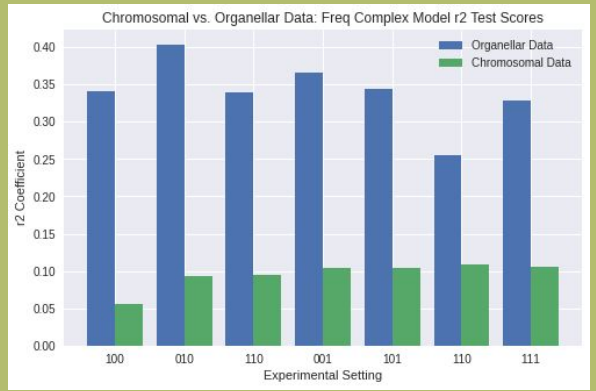
r2
Simple

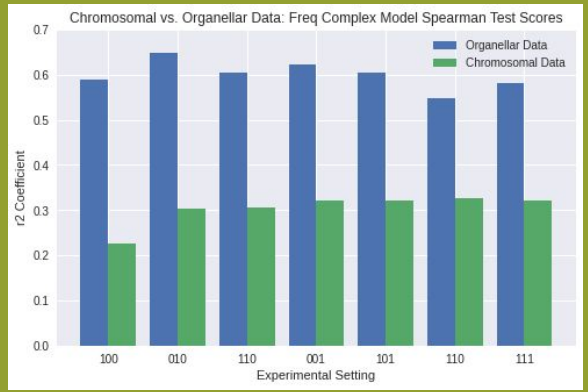




Spearman
Simple

r2
Complex

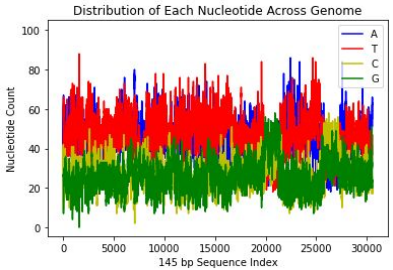


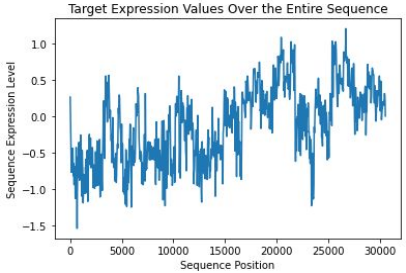


Spearman
Complex

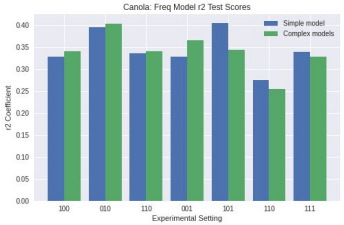
Organelle vs. Chromosomal Data

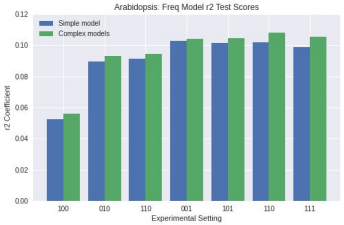
- Why is it that everything performs better on the organelle data?
  - Organelles are smaller and simpler (easier to pick out patterns)
  - Nucleotide content may matter more than longer sequences





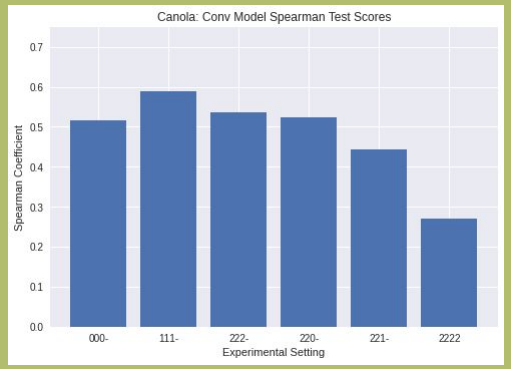
- Additionally, there seemed to be more randomness in the results of the organelle data

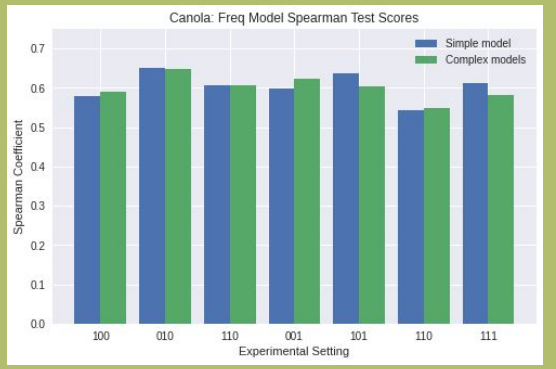




Frequency vs. Conv Models

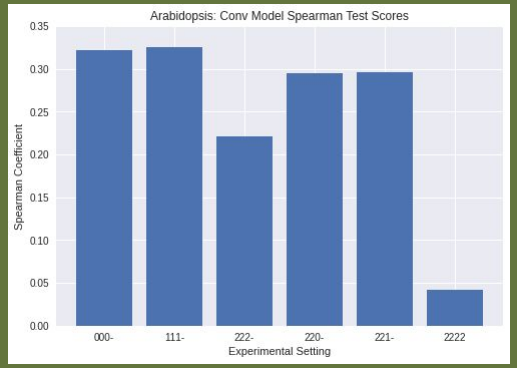
When using organellar data, the frequency models appear to match or outperform the conv models

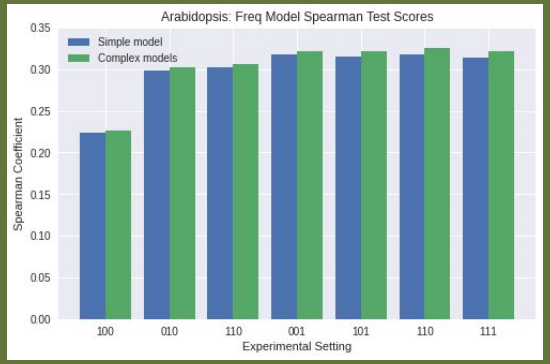




Using convolutional models with chromosomal data gave the best score, but not by much

Overall, the scores look similar



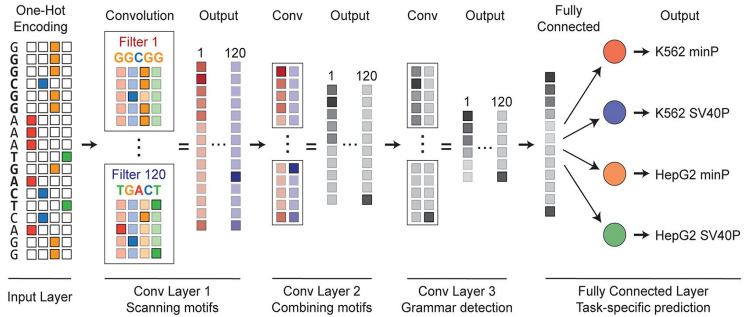


Comparison to MPRA-DragoNN Results

- Recall: Arabidopsis chromosomal data best scores
  - r2: 0.108 (freq, complex, `110`)
  - Spearman: 0.325 (conv, `111-`)

- MPRA-DragoNN reported a conv model Spearman score of 0.28

- These are not directly comparable (different organism, slightly different model)
  - However, given our results, it is possible that the authors did not need a complex convolutional model at all to achieve their results



Limitations

- Our analysis is only as good as the data is
  - Multiple arbitrary decisions made so the analysis could proceed
  - This work is preliminary, as we do not have protein crop data yet
- Comparing cross-organism and cross-organelle currently
- Computational power and data size
  - Could not utilize all 24 million rows of chromosomal data
  - Trade-off between data size and computational time
- Further work would need to be done to see if these predictions are viable *in vitro*



Future Work

Plans for the next two years

What next?

1. Test new techniques for prediction:
    a. Recurrent / attention based models
    b. Ensemble learning
    c. Neural architecture search
    d. Incorporating multiple organisms for generalizability
    e. Predict up/down regulated instead of a real number

2. New data
    a. NRC is completing more STARR-Seq experiments with new organisms
        i. Focus on protein crops

3. Interpretation
    a. Which patterns of sequences in particular seem important for expression enhancement?
    b. Test methods like DeepLIFT

Acknowledgements

Shazan Jabbar
Lead ML Scientist

Mara Cairo
Product Owner

Daniel Erin
Grant Writer

David Konkin
Principal Investigator

Dustin Cram
Bioinformatician

Morgan Kirzinger
Bioinformatician



We need more collaborators!

If this project seems interesting to you, feel free to reach out to:

- Myself: sarah.davis@amii.ca or sdavis1@ualberta.ca
- Mara Cairo: mara.cairo@amii.ca
- Shazan Jabbar: shazan@amii.ca

This is a two-year project, and we will officially looking for another intern for the 2022 winter term.


Thank you!
Questions?