ZIMIN FOUNDATION

SMTB

Generating Synthetic Data

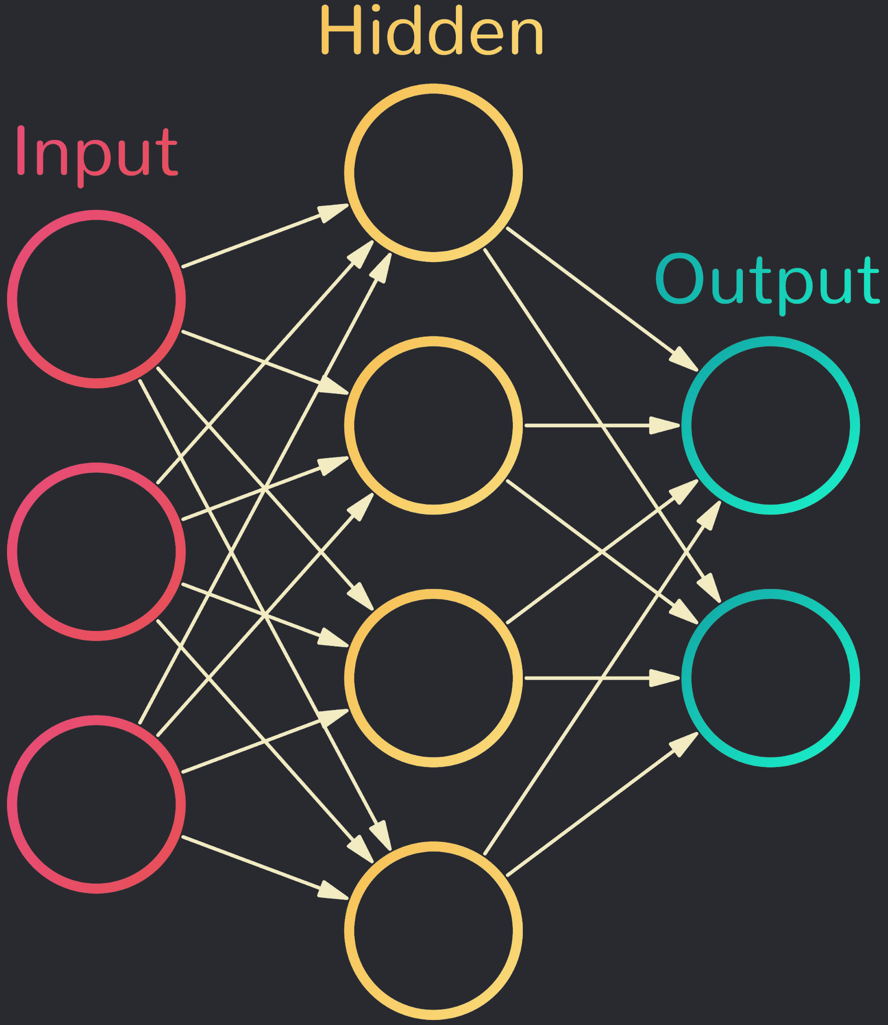
Rauf Verdiev, Roman Kotovich, Vera Terenteva, Elizaveta Terekhova, Laura Avinyó

Project goal

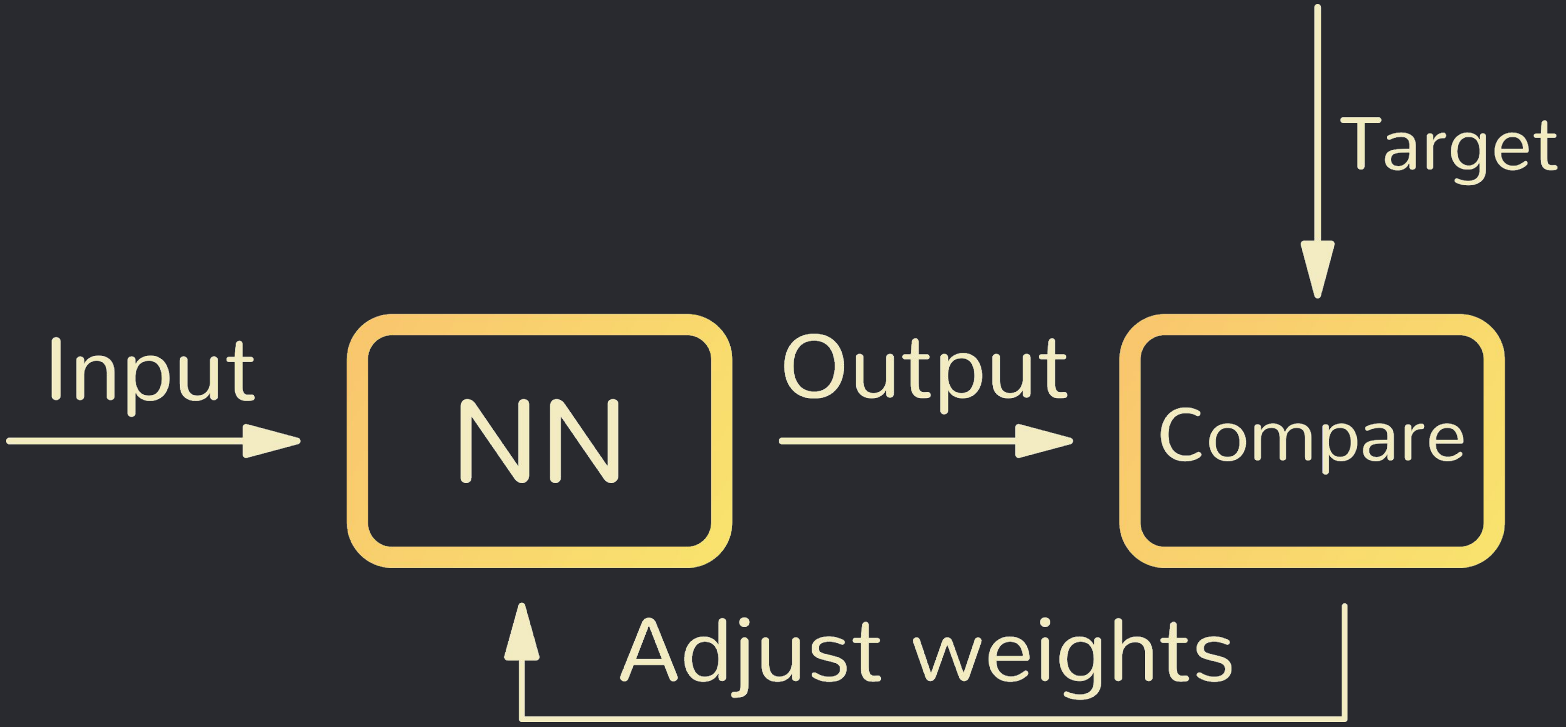
- Generate synthetic data
    - HMM
    - GANs
- Analyze generated sequences
    - Zm
    - Decision tree
    - Shannon Entropy

Why did we do this?

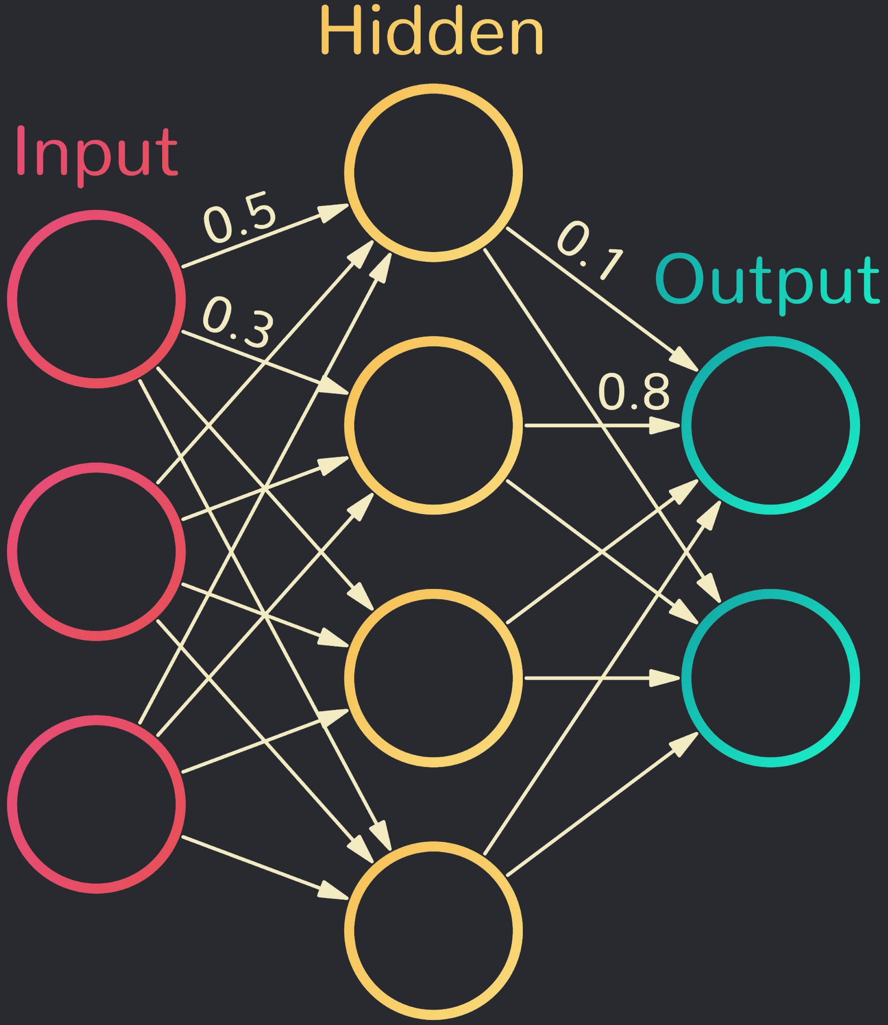
Neural Networks

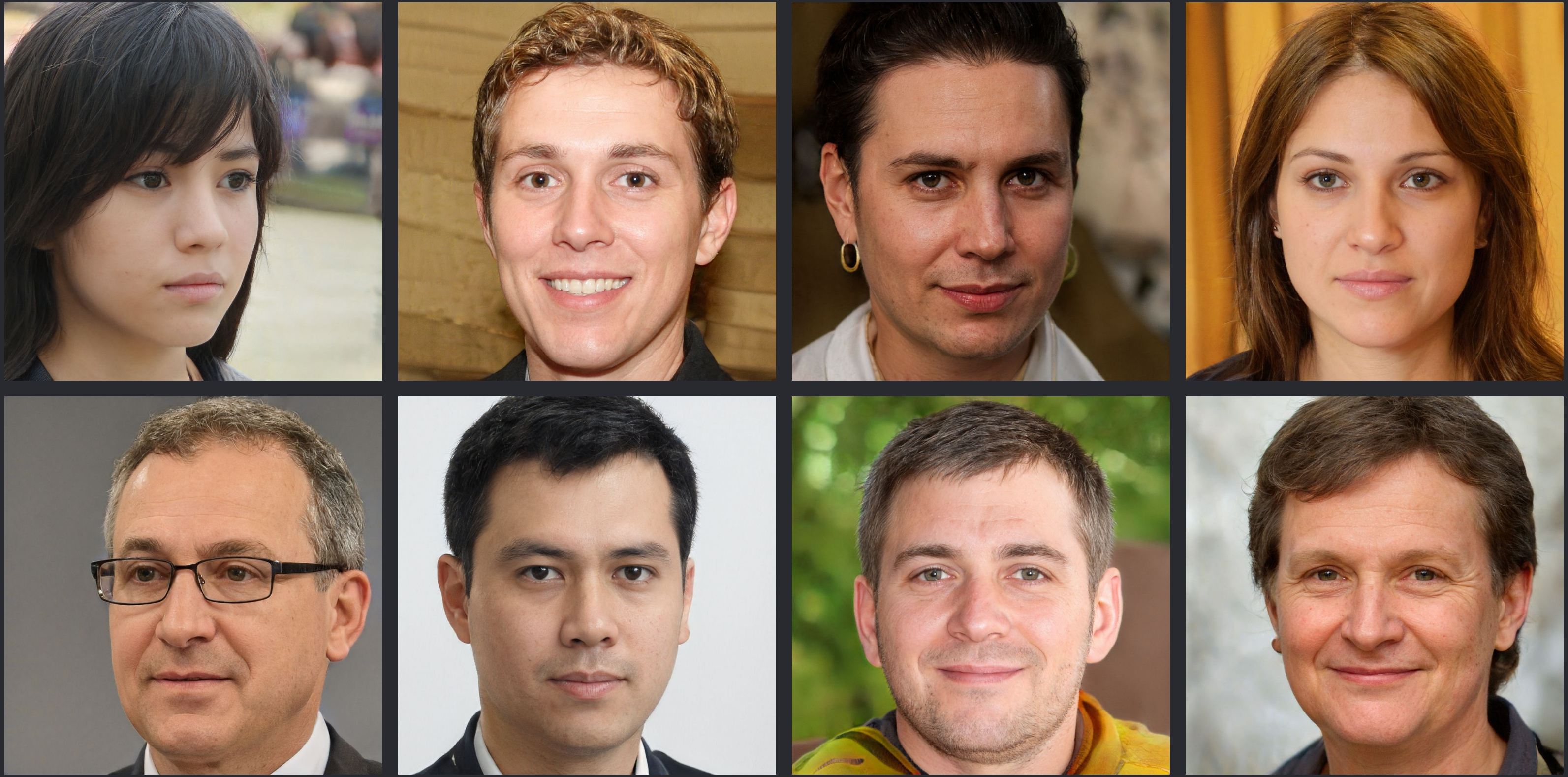


Training



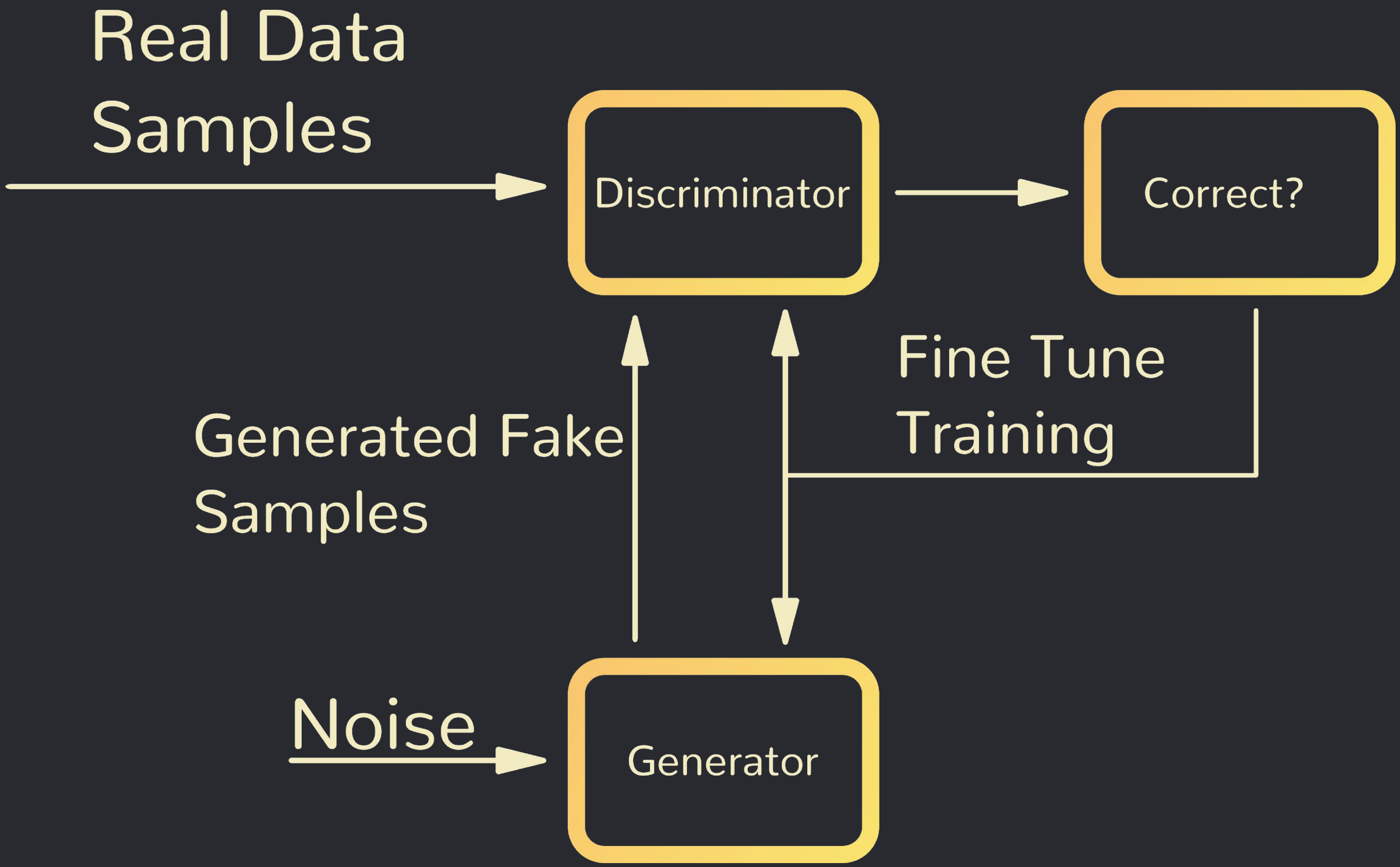
Trained Network





GANs

Generative
Adversarial
Neural Networks



Real Data Samples

Discriminator

Correct?

Fine Tune Training

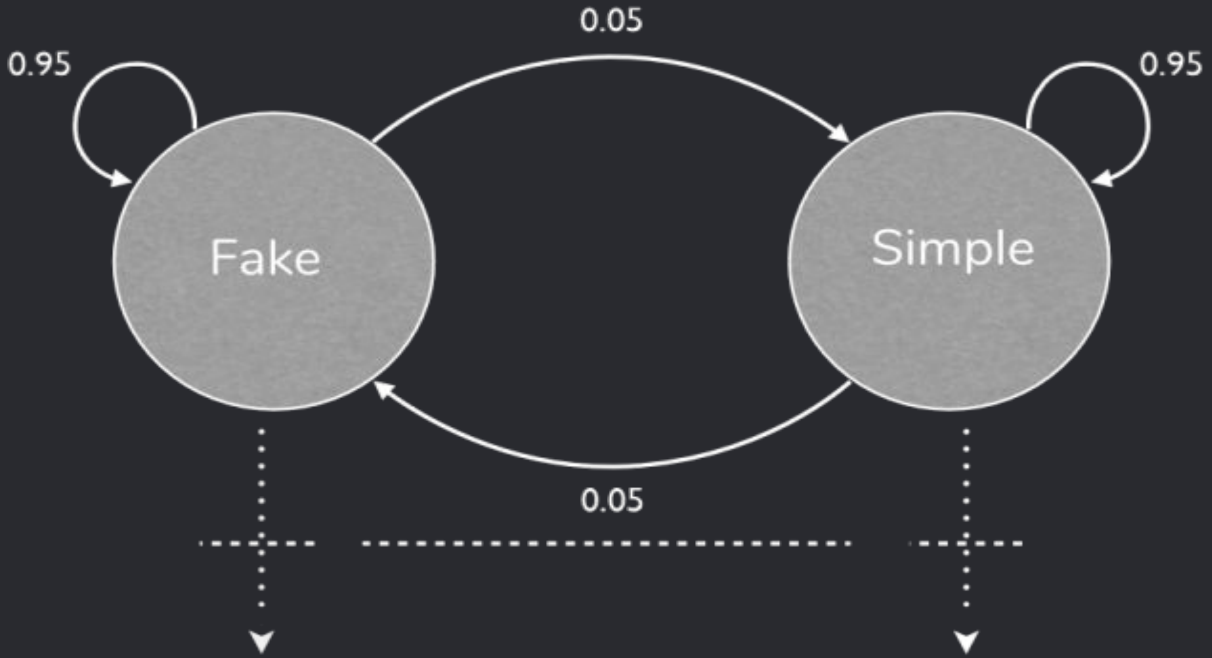
Generated Fake Samples

Noise

Generator

Hidden Markov Models (HMM)

Hidden states



Observed states

P(1|Fake)=⅙
…
P(5|Fake)=⅙
P(6|Fake)=½

P(1|Simple)=⅙
...
P(6|Simple)=⅙

Hidden Markov Models (HMM)

1 2 4 5 1 3 2 4 5 6 3 2 4 P = E(1|T)*T(T|T)*E(2|T)*T(T|T)*…*E(3|T)*T(T|F)*E(2|F)*T(F|F)*…

3 1 5 2 5 6 3 2 4 1 5 6 6 P = E(3|T)*T(T|F)*E(1|F)*T(F|T)*E(5|T)*T(T|F)*E(2|F)*T(F|T)*…

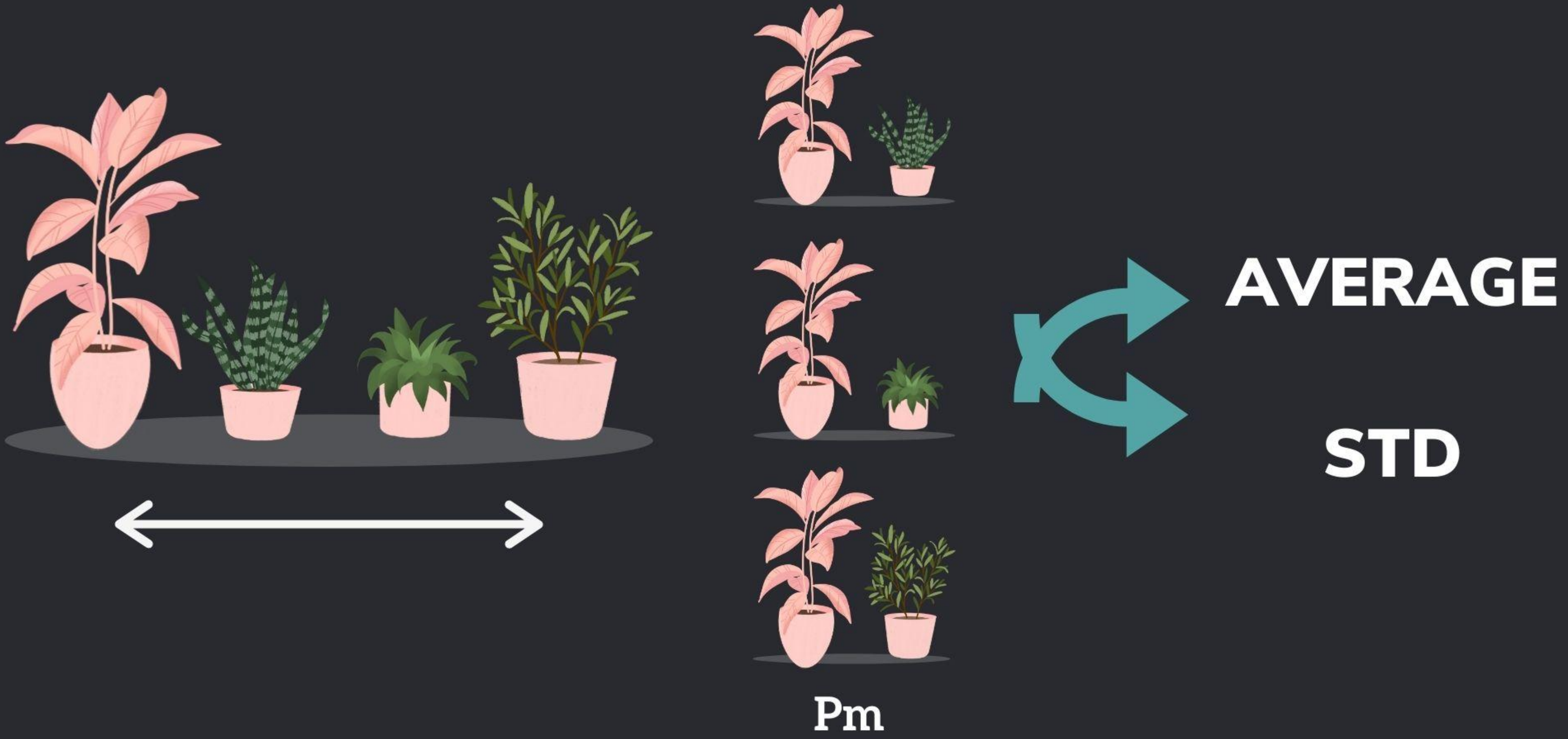
We can generate some sequences!

Emission match = [1: {A=0.01; B=0.06; …}, 2:{A=0.07; B=0.05; …}, …, 39:{A=0.01; B=0.09; …}]

Insertion match = [1: {A=0.03; B=0.05; …}, 2:{A=0.04; B=0.06; …}, …, 39:{A=0.07; B=0.05; …}]

Transition = [1: { m->m: 0.92; m->i: 0.10; m->d: 0.11; i->m: 0.70; i->i: 0.20; d->m: 0.90; d->d: 0.01}, …]

HOW WE CHECK DATA?



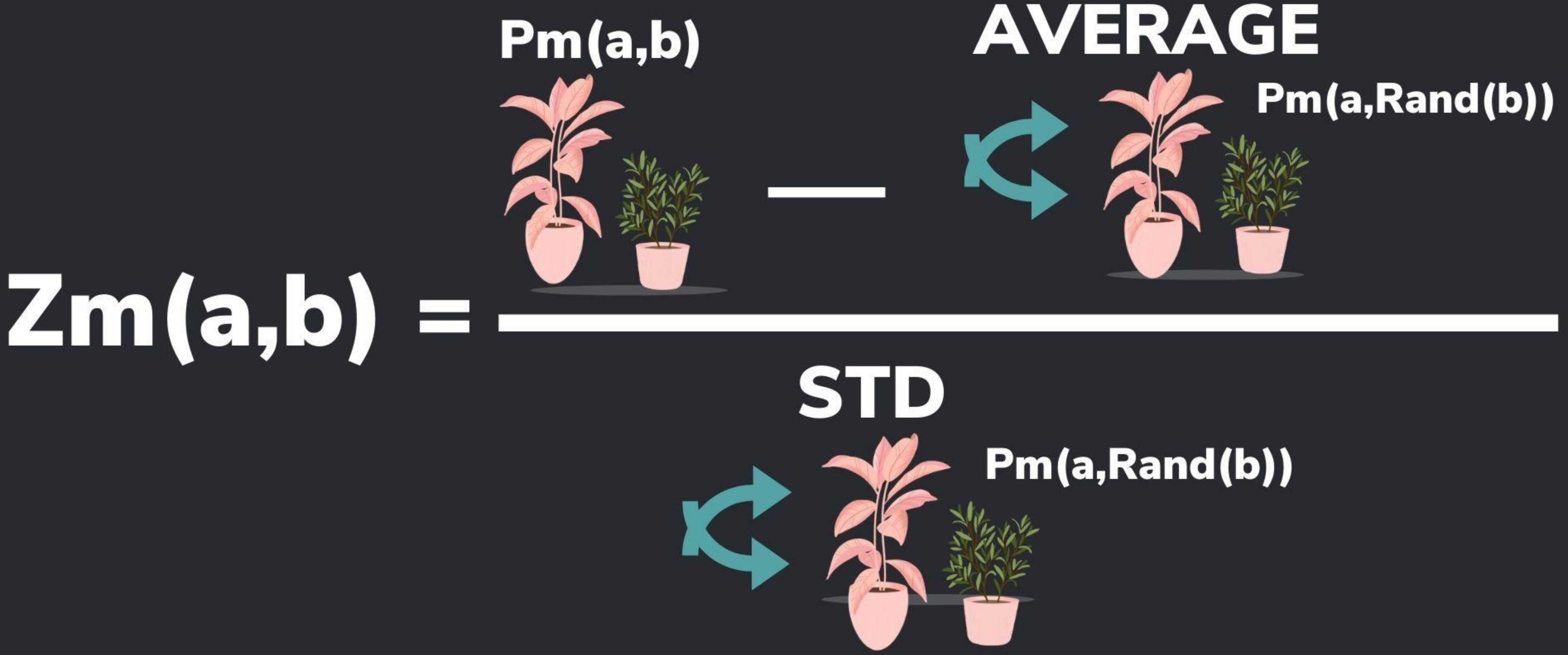
Pm

AVERAGE

STD

HOW WE CHECK DATA?

Zm associate score



Zm(a,b) =

Pm(a,b)

—

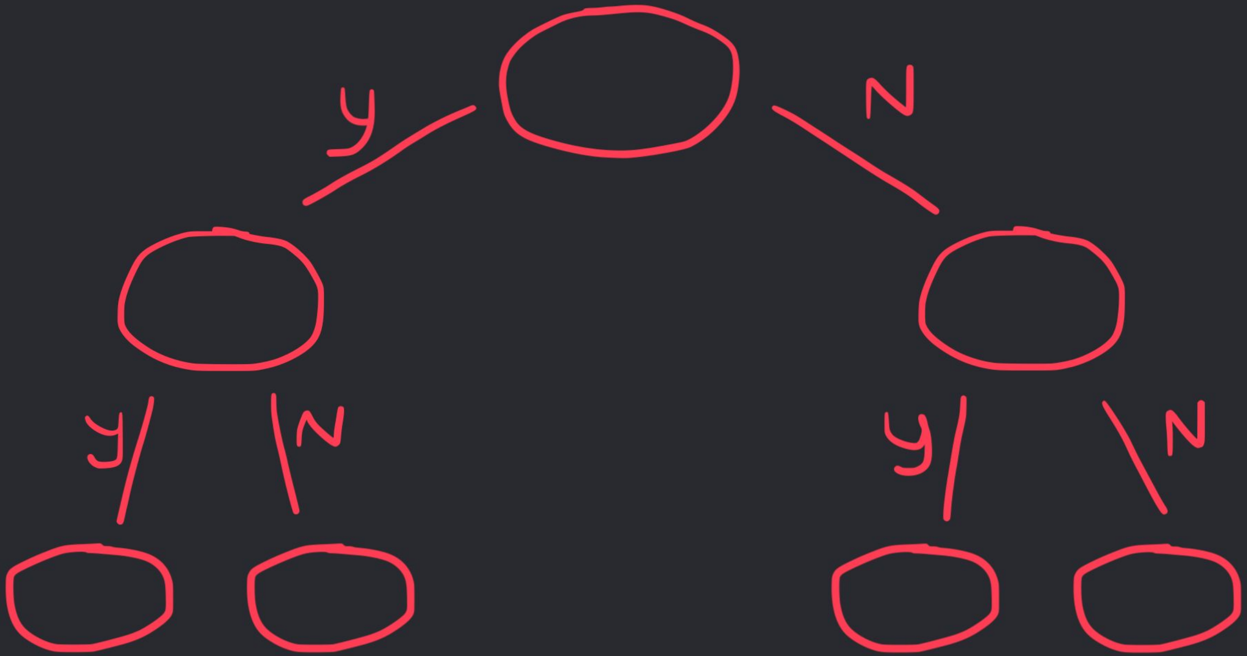
AVERAGE

Pm(a,Rand(b))

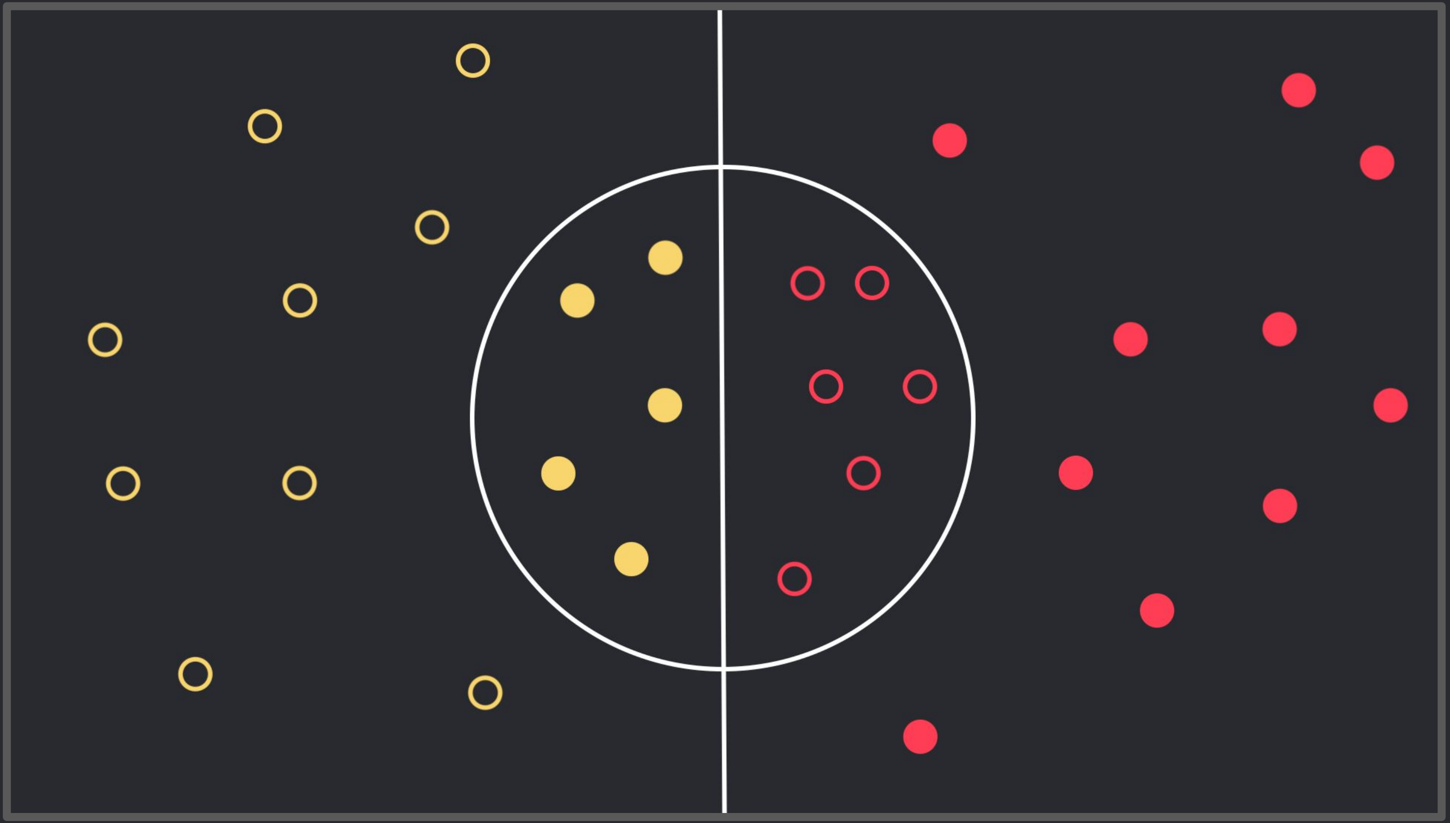
STD

Pm(a,Rand(b))

How to grow a decision tree?



Accuracy, Precision and Recall



Shannon Entropy

The Shannon entropy can measure the uncertainty of a random process

Its lower values imply less uncertainty

For each column of multiple-sequence alignment (MSA) Shannon is calculated as:

$$s.e. = -\sum_{i=1}^{20} p(x_i) log_{20} p(x_i)$$

$p(x_i)$ - is the probability to find the amino acid *i* at the column of MSA

In terms of columns in MSA its low values represent highly conservative positions, whereas high entropy shows diversity at certain position

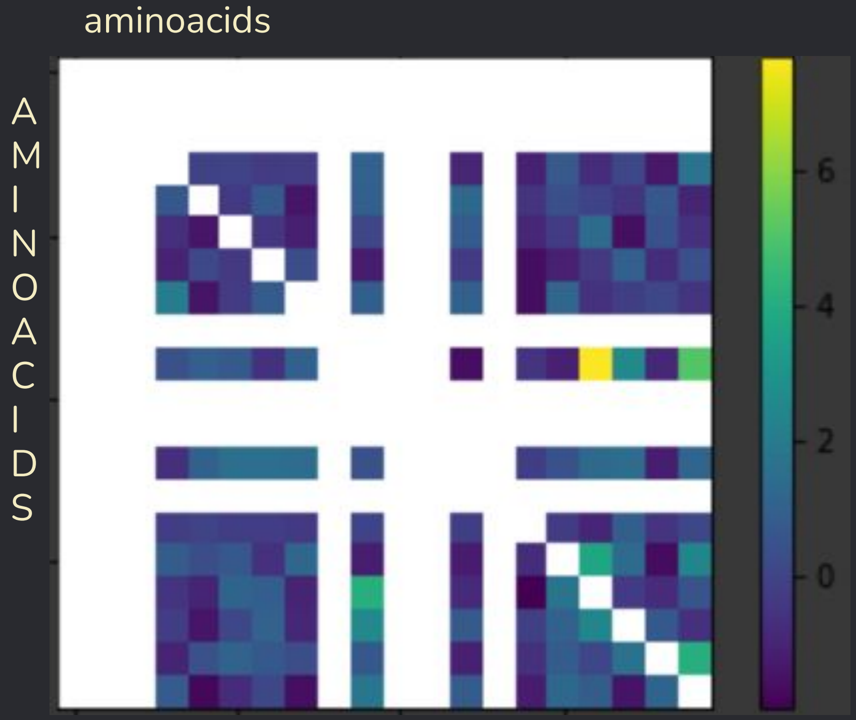
Project goal

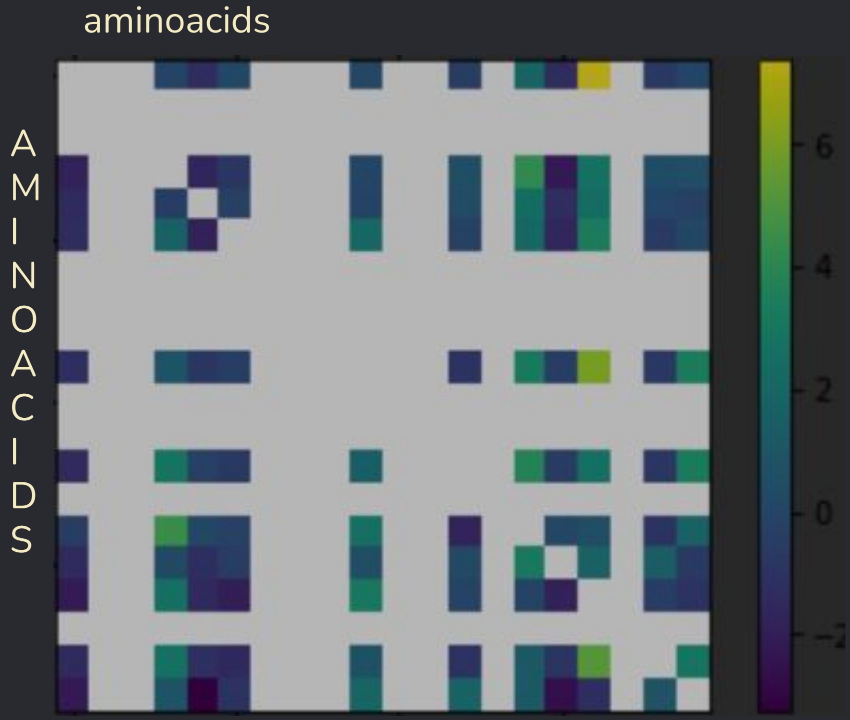
1. Generate Synthetic data with the different methods
    a. GANs
    b. HMM
    c. Random Baseline
2. Test how good or bad the data is:
    a. ML discriminator -> Overall quality
    b. Shannon’s Entropy -> Functional
    c. Zm -> 3D structure

Why did we do this?

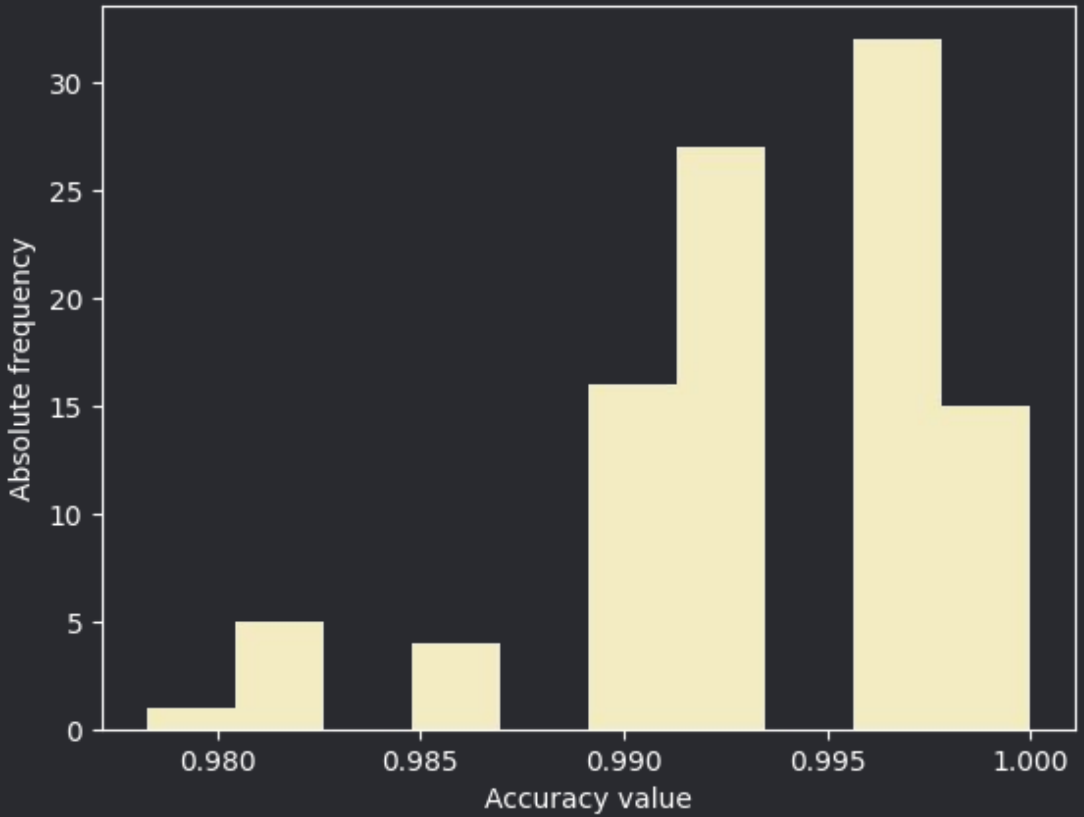
Results

Zm results on generated and real data

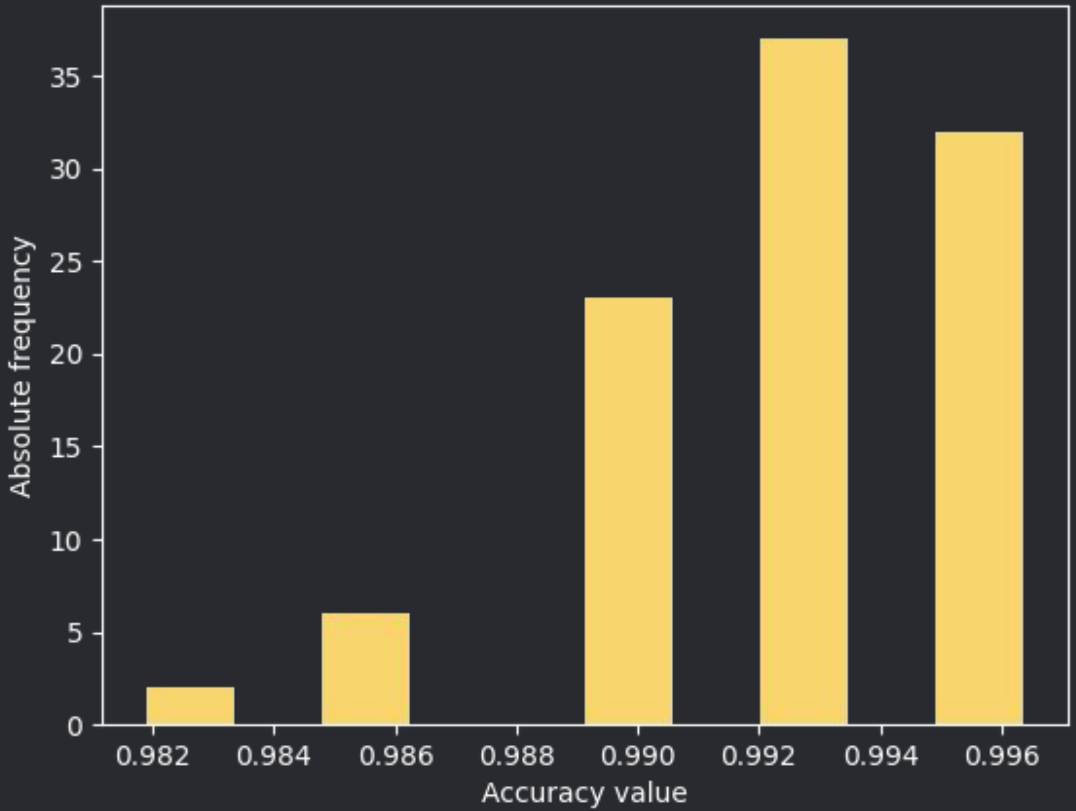




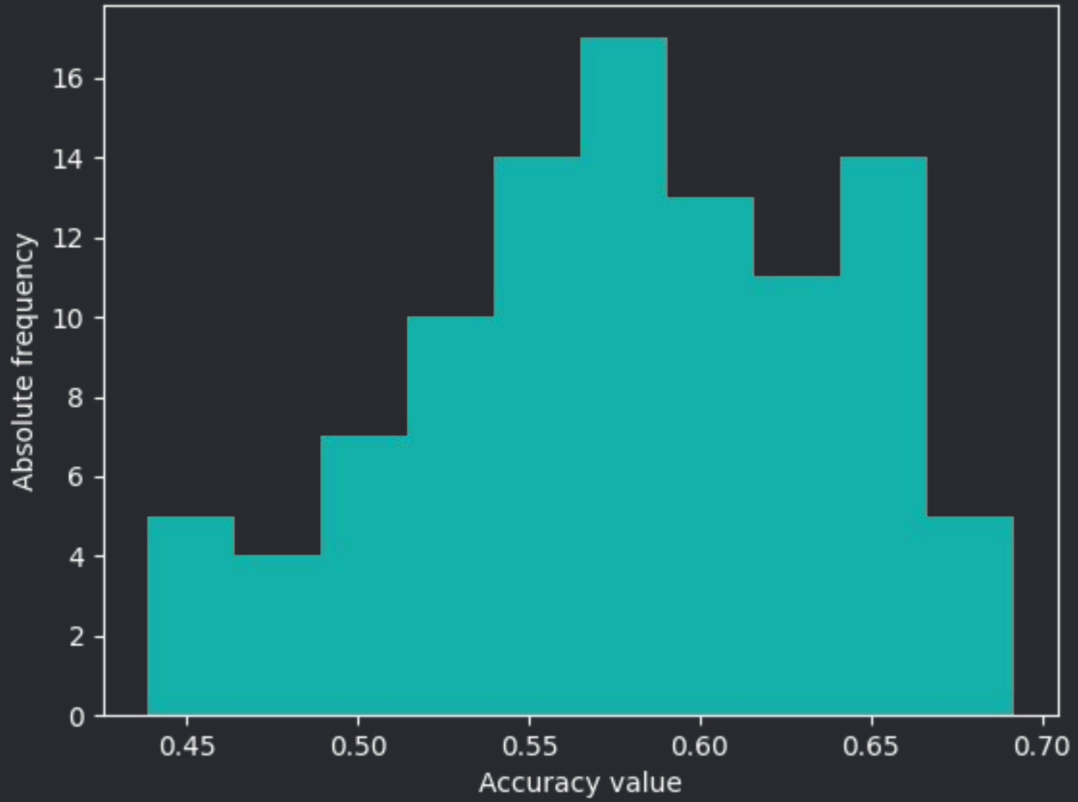
Accuracy - baseline



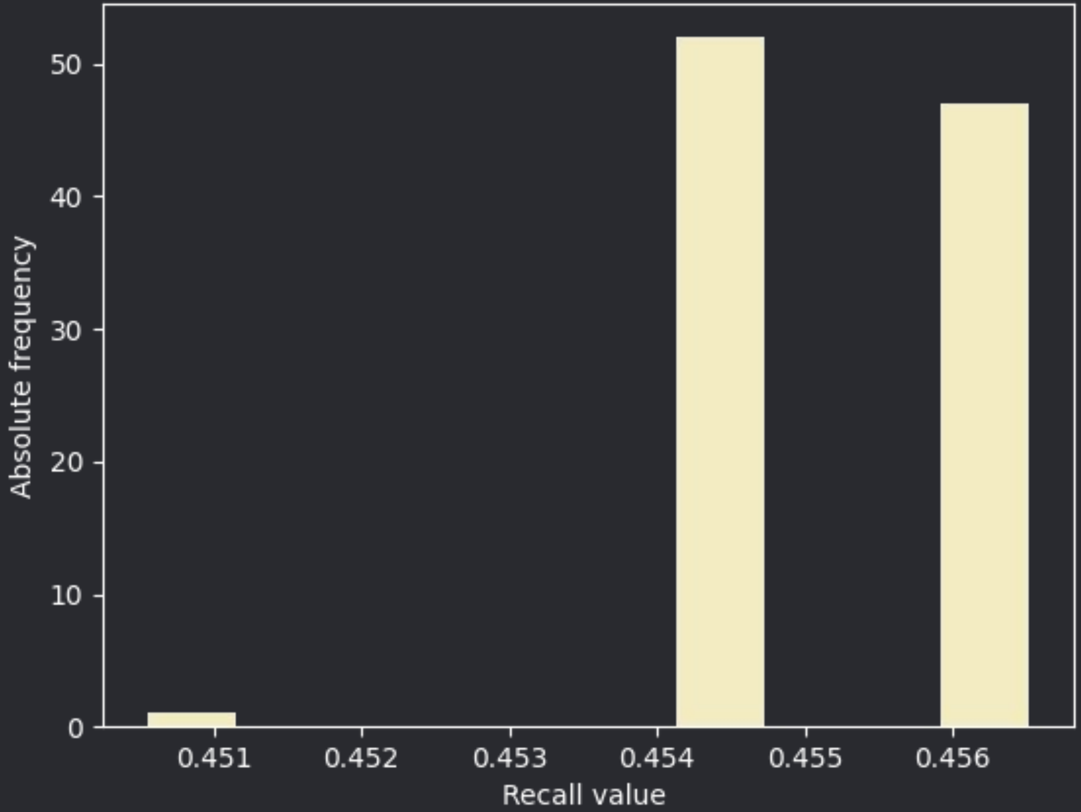
Accuracy - GAN



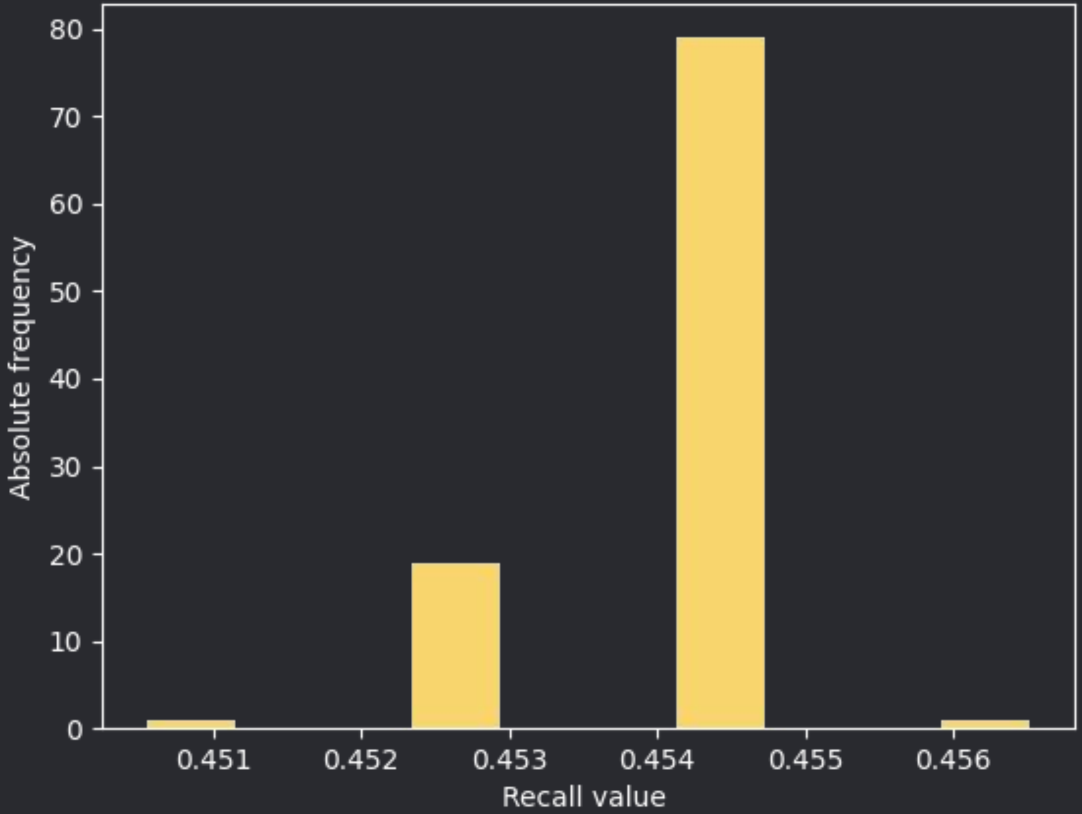
Accuracy - HMM



Recall - baseline

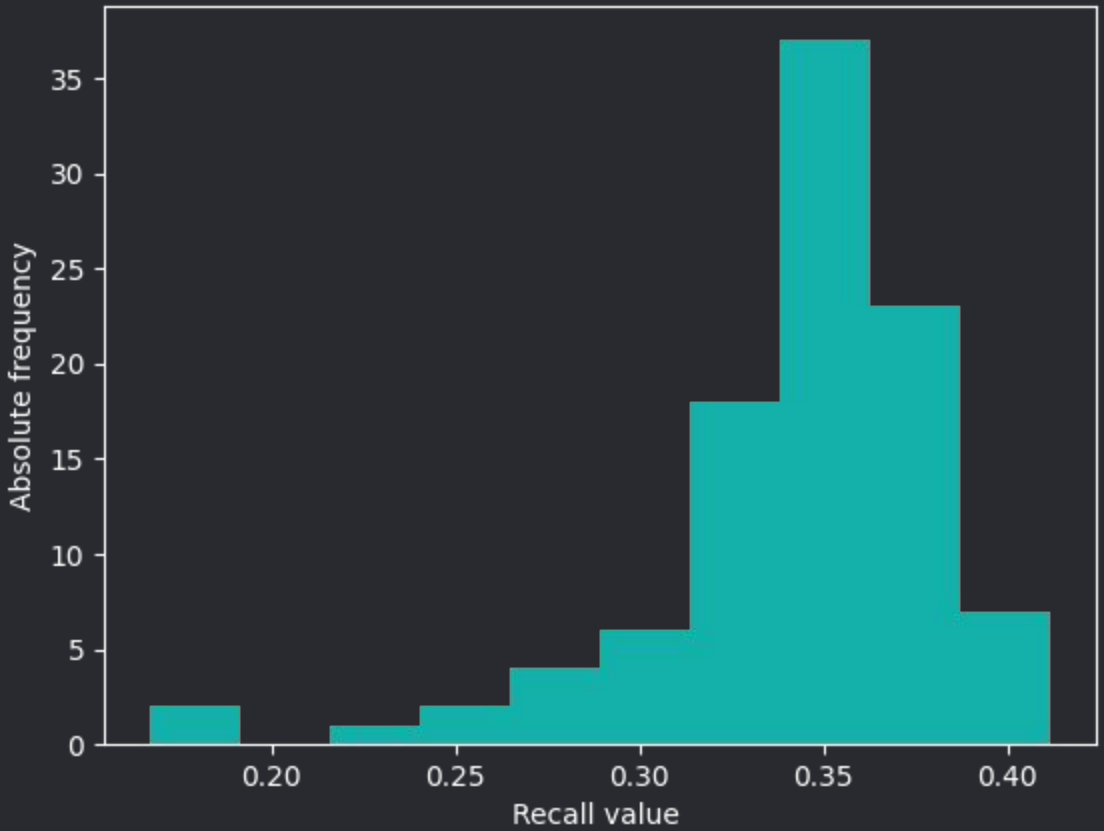


Recall - GAN

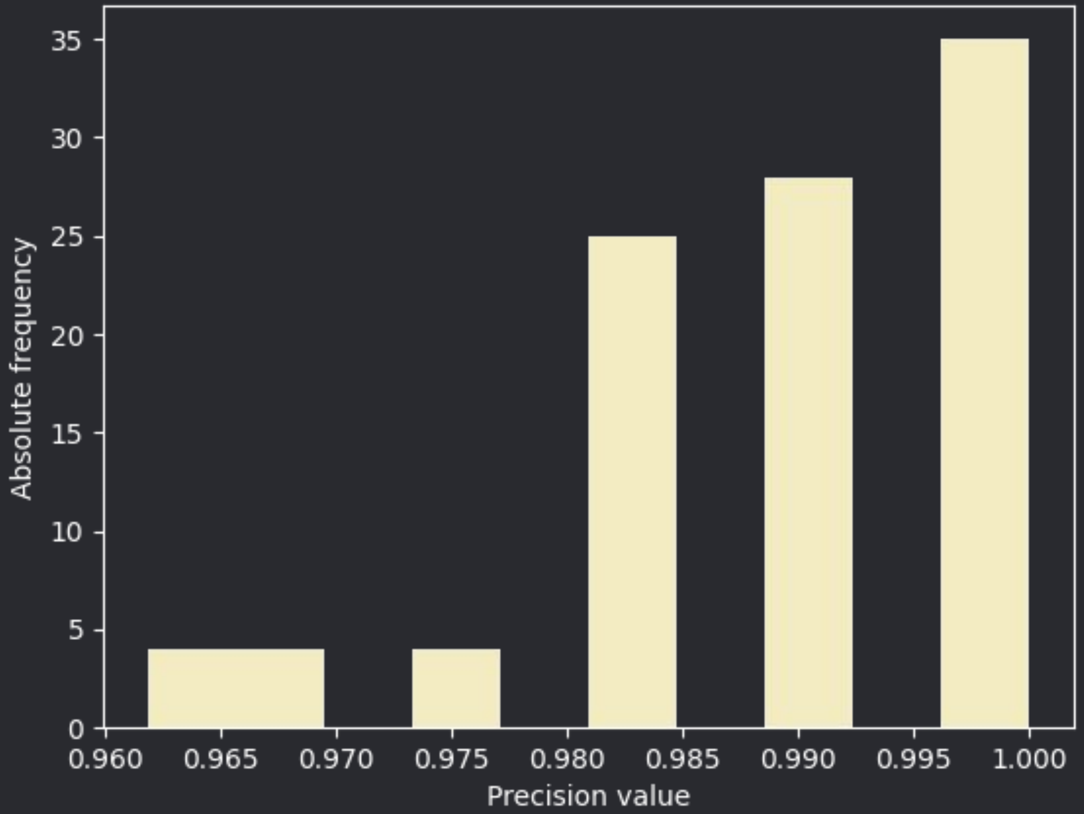


Absolute frequency
0
10
20
30
40
50
60
70
80
0.451
0.452
0.453
0.454
0.455
0.456
Recall value

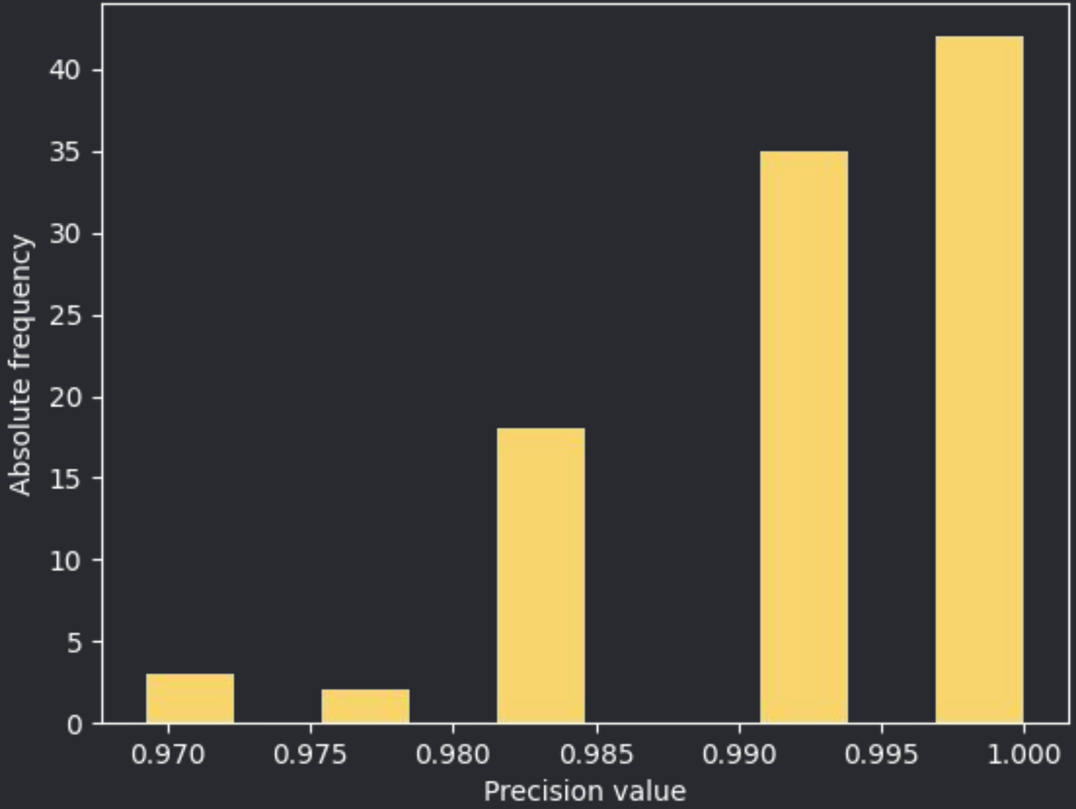
Recall - HMM


Absolute frequency
0
5
10
15
20
25
30
35
0.20
0.25
0.30
0.35
0.40
Recall value

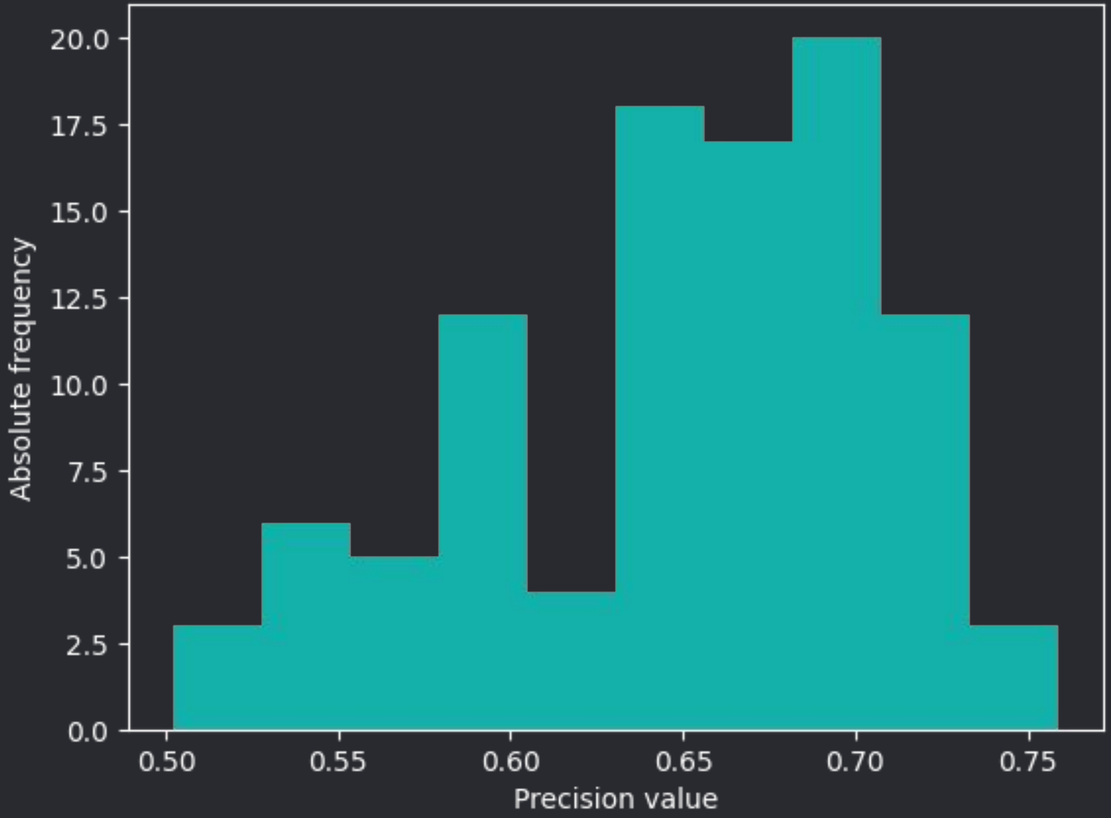
Precision - baseline



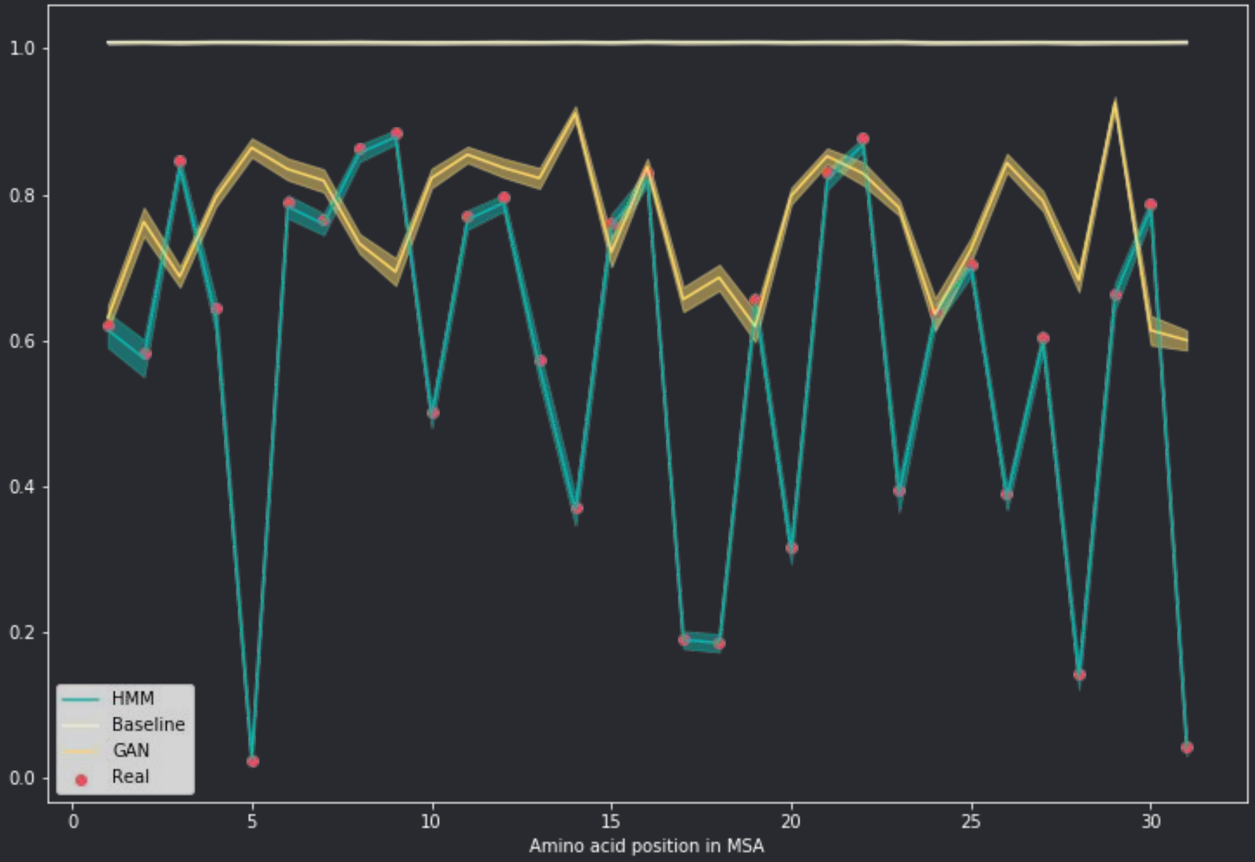
Precision - GAN



Precision - HMM



Shannon Entropy



So did we do well?