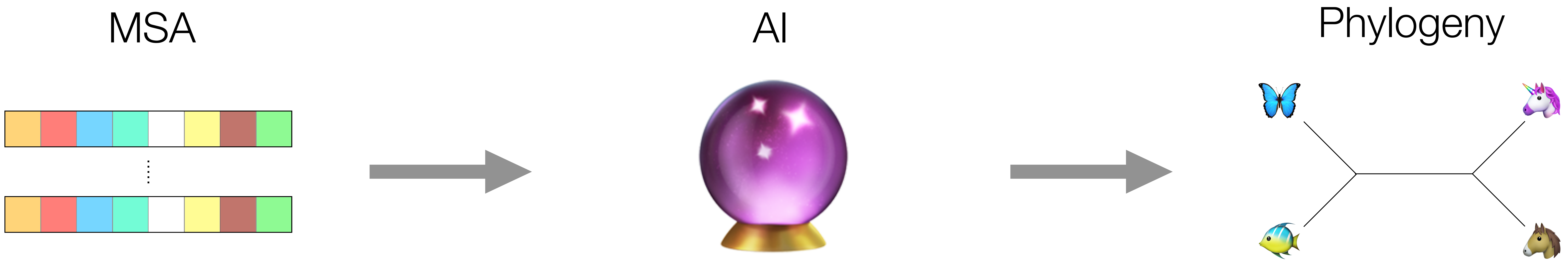
Simulations of Sequence Evolution

How (Un)realistic They Are and Why

J. Trost, **J. Haag**, D. Höhler, L. Jacob, A. Stamatakis & B. Boussau

Motivation

Motivation

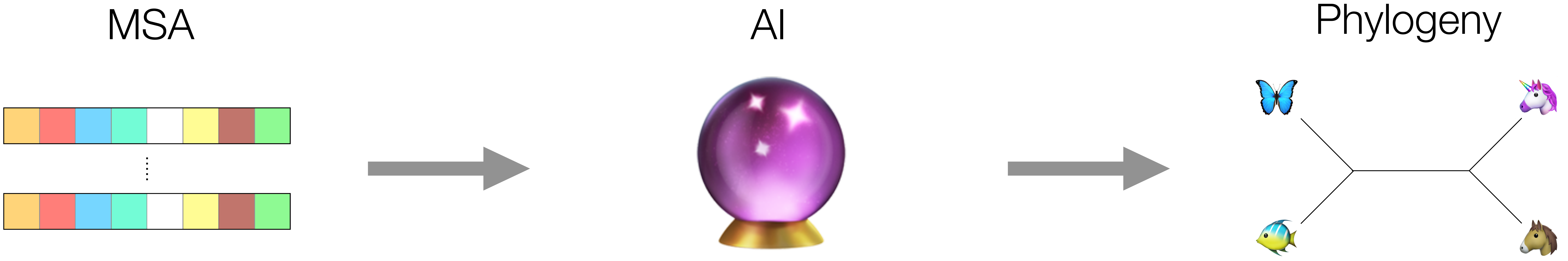


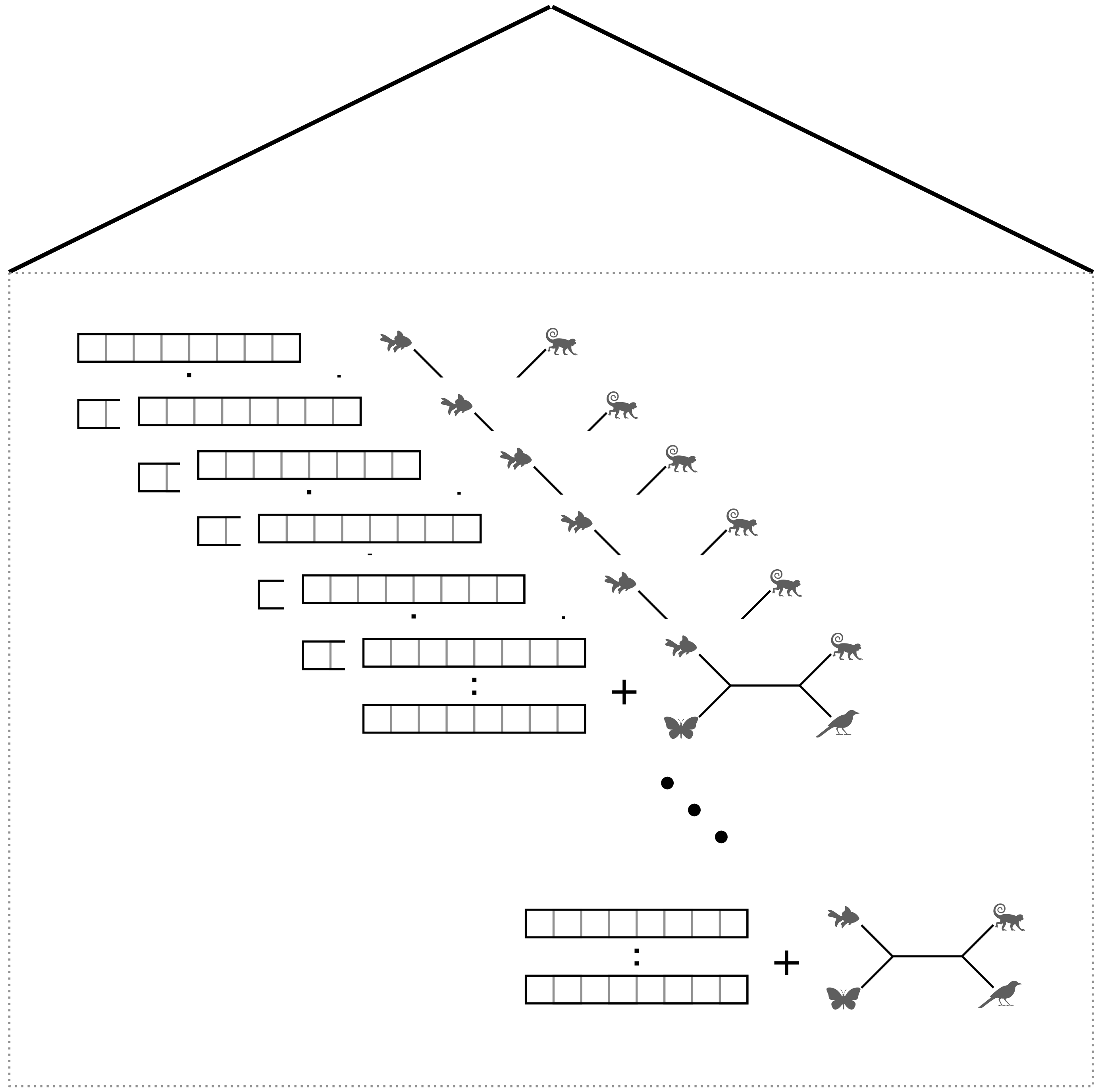
MSA

AI

Phylogeny

Motivation



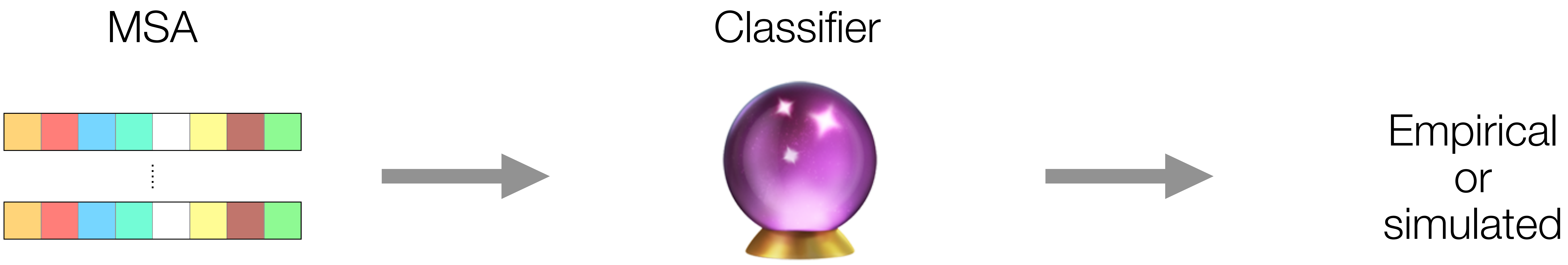


Simulations?

Simulated data == empirical data?

How (un)realistic are simulations?

Overview

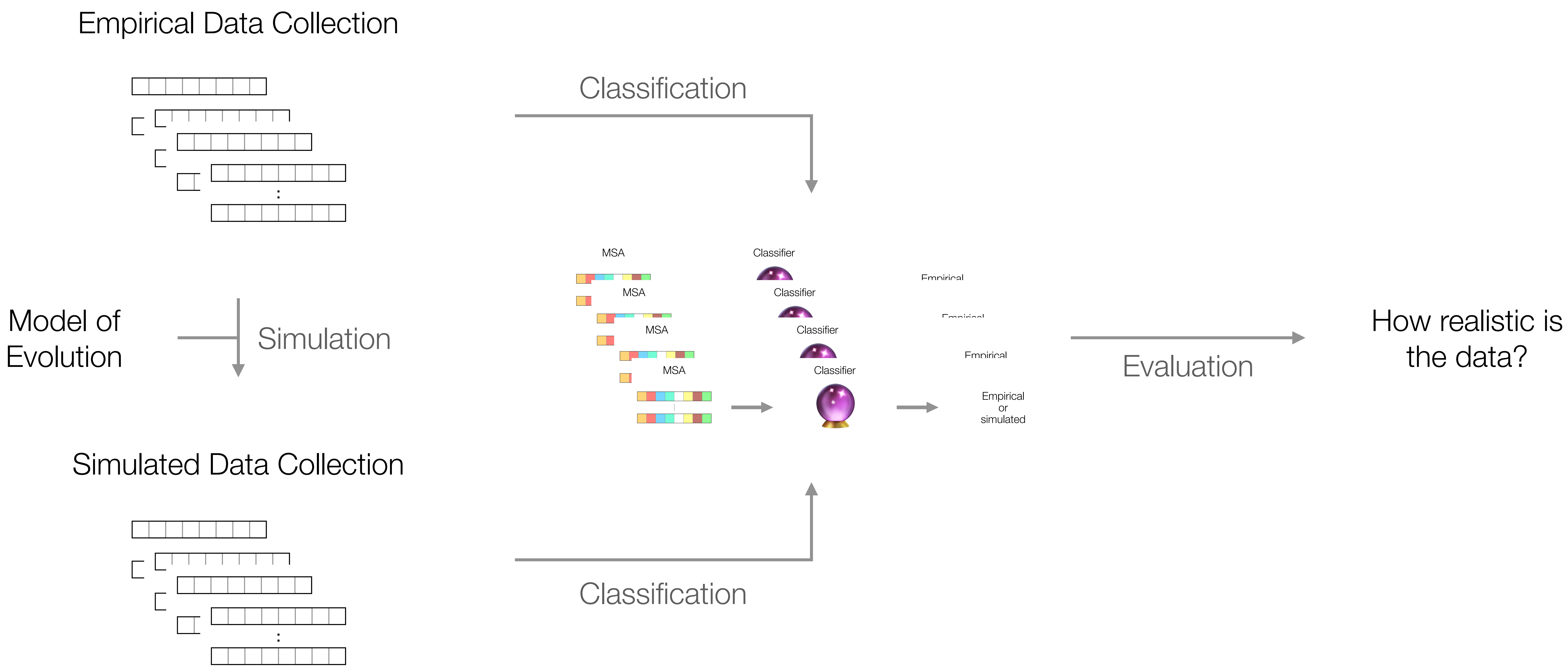


MSA

Classifier

Empirical
or
simulated

Overview



Empirical Data Collection

Classification

Model of Evolution

Simulation

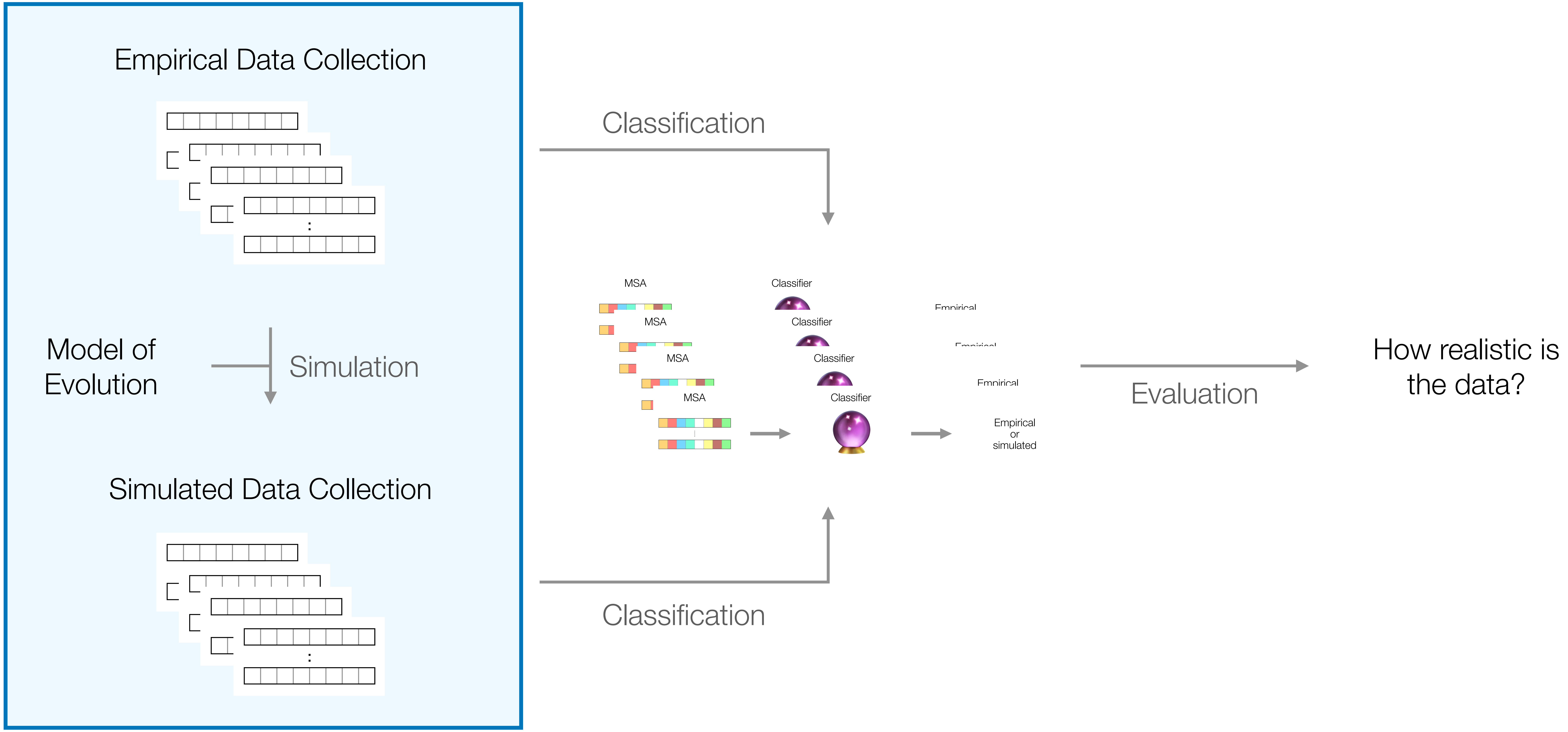
Simulated Data Collection

Classification

Evaluation

How realistic is the data?

Overview



Empirical Data Collection

Model of Evolution

Simulation

Simulated Data Collection

Classification

Classification

Evaluation

How realistic is the data?

Alignment Simulation

- 15 simulated data collections
  - DNA: 5 + 2
  - Protein: 7 + 1
- Models of Evolution:
  - DNA: JC, HKY, GTR, GTR+G, GTR+G+I
  - Protein: Poisson, WAG, LG, LG+C60, LG+S256, LG+S256+G4, LG+S256+GC
- Empirical data collections
  - DNA: TreeBASE (9460 MSAs)
  - Protein: HOGENOM (6971 MSAs)

Alignment Simulation

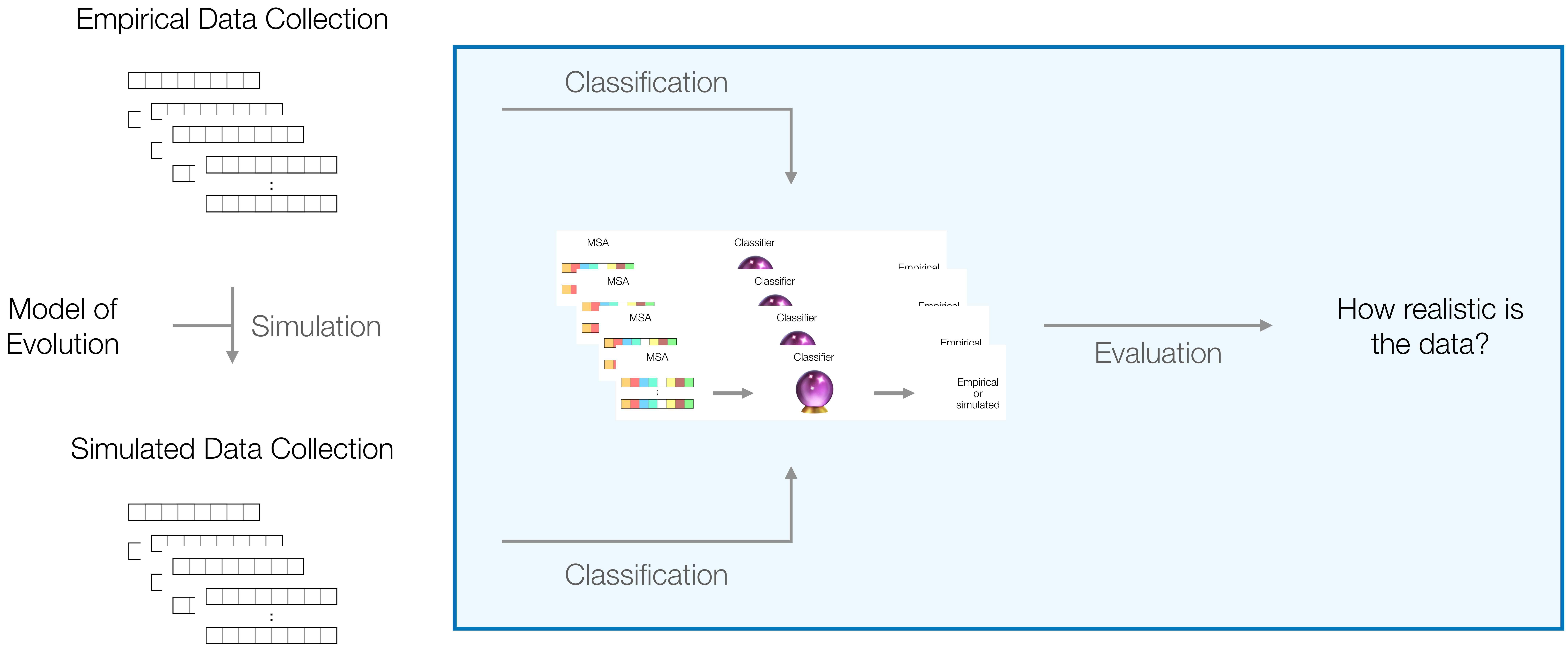
- Phylogeny + simulation parameters based on empirical data collections
- Simulation Tool: AliSim
- Simulations without indels
- Indel Simulation:
    - Mimick approach: superimpose gap patterns
    - SPARTA approach: empirical indel parameteres (SpartaABC)

Simulated Data Collections

- DNA (5 + 2):
  - Gapless: JC, HKY, GTR, GTR+G, GTR+G+I
  - With Indels: GTR+G+I+mimick, GTR+G+I+sparta

- Protein (7 + 1):
  - Gapless: Poisson, WAG, LG, LG+C60, LG+S256, LG+S256+G4, LG+S256+GC
  - With Indels: LG+S256+GC+sparta

Overview



Empirical Data Collection

Model of Evolution

Simulation

Simulated Data Collection

Classification

Evaluation

How realistic is the data?

Classification

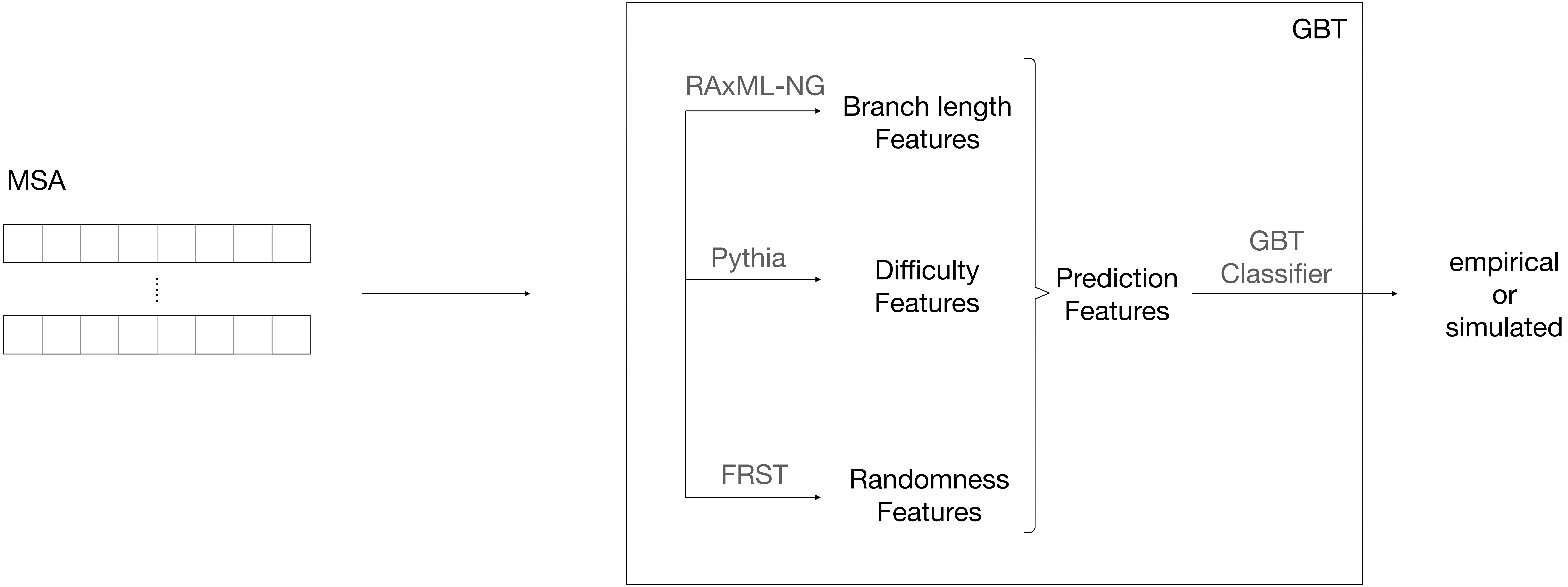
Training and Evaluation

- Two distinct classifiers
  - Gradient Boosted Trees (GBT)
  - Convolutional Neural Network (CNN)
- 1 classifier each for each simulated data collection

  ⇒ 15 GBTs + 15 CNNs

- Training data: simulated + empirical data collection
- 10-fold CV + Balanced Accuracy (BACC)
- Final accuracy: average BACC over all 10 folds

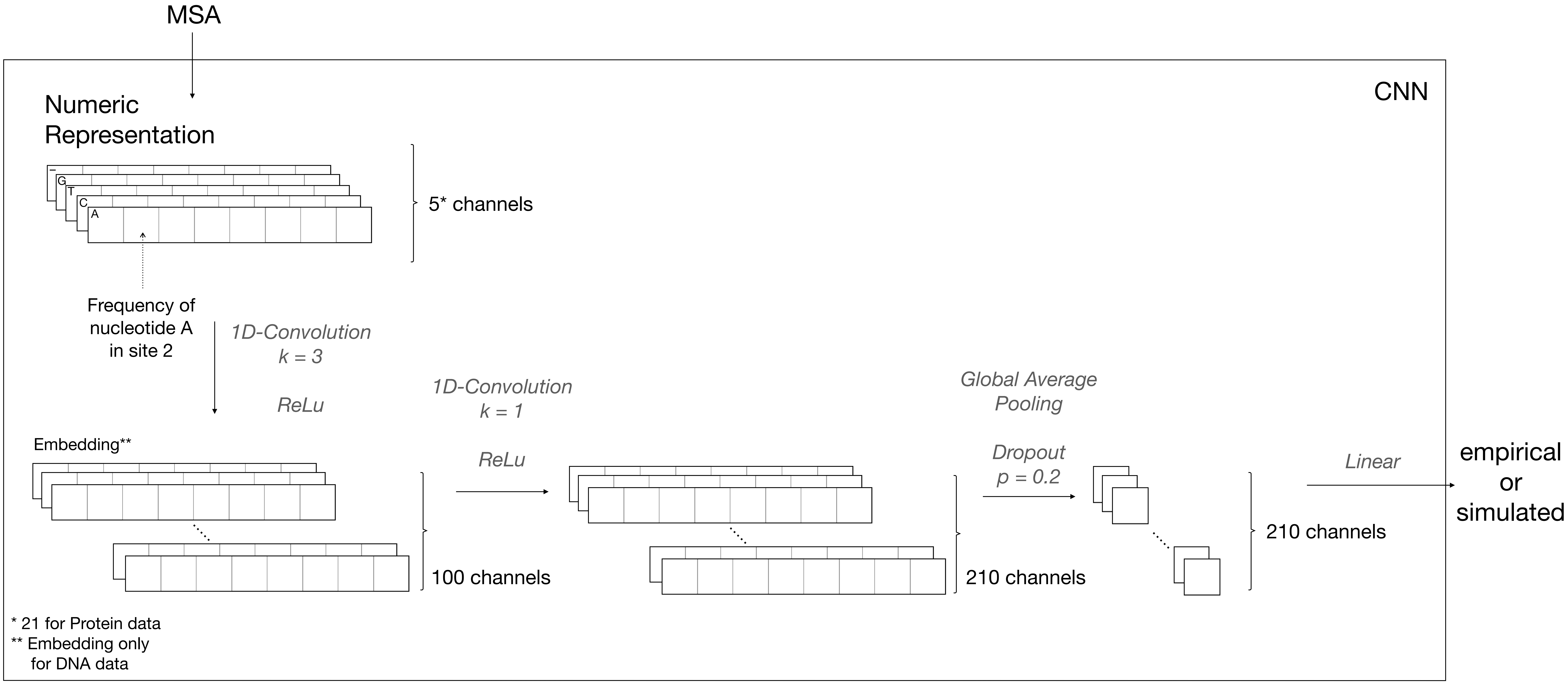
Gradient Boosted Trees



GBT: Features

- Branch length features
  - Based on RAxML-NG tree inference
  - Average branch length, maximum branch length, …
- Difficulty features (Pythia)
  - Based on Pythia difficulty prediction
  - Predicted difficulty, sites-per-taxa ratio, proportion of invariant sites, …
- Randomness features (FRST)
  - Based on parsimony substation counts and FRST
  - Entropy, Serial Correlation Coefficient, …

Convolutional Neural Network



* 21 for Protein data

** Embedding only for DNA data

Classification Performance

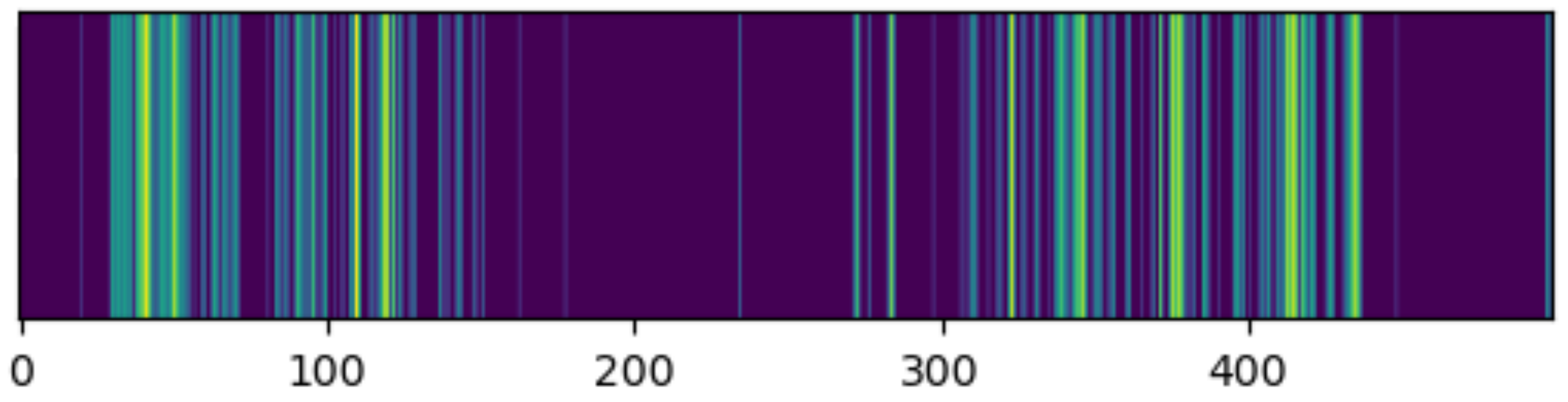
	BACC	
	GBT	**CNN**
DNA data collections		
JC	0.96	0.99
HKY	0.96	0.99
GTR	0.94	0.93
GTR+G	0.89	0.94
GTR+G+I	0.89	0.94
GTR+G+I+mimick	0.77	0.97
GTR+G+I+sparta	0.94	0.97

	BACC	
	GBT	**CNN**
Protein data collections		
Poisson	0.99	0.9996
WAG	0.99	0.97
LG	0.99	0.95
LG+C60	0.98	0.99
LG+S256	0.99	0.995
LG+S256+G4	0.99	0.99
LG+S256+GC	0.98	0.99
LG+S256+GC+sparta	0.99	0.996

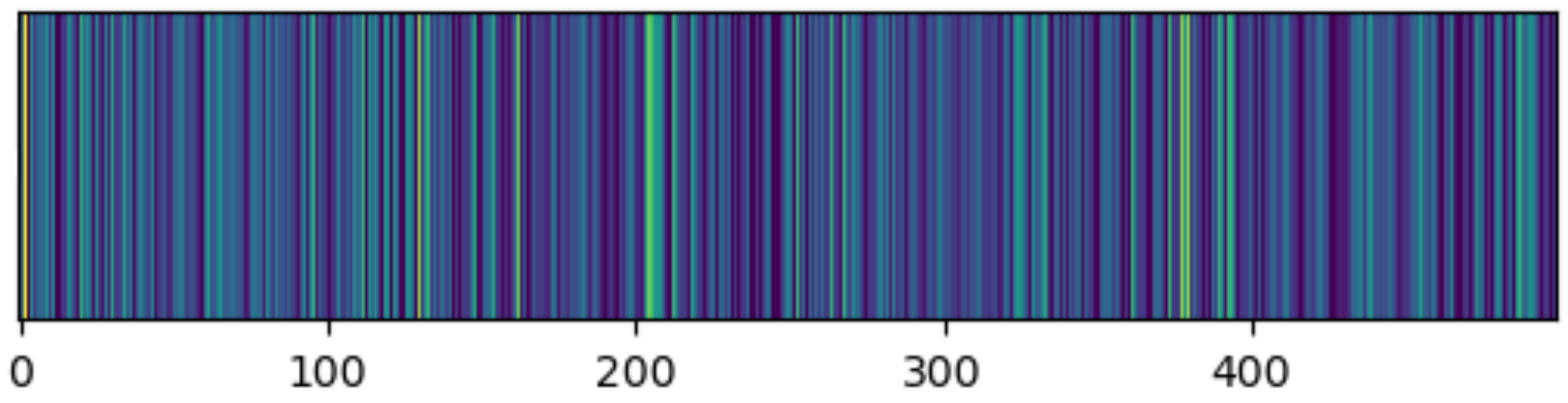
GBT Feature Importance

- Most important features: randomness metrics

Empirical DNA MSA



Simulated DNA MSA



GBT Feature Importance

- Most important features: randomness metrics
  - Randomness across sites
  - Randomness within sites
- Empirical data:
  - Higher proportion of invariant sites
  - Longer branches

CNN: Feature Importance

- Logistic Regression
  - Feature: Alignment-wise AA/Nucleotide frequency
- DNA data:
  - BACC ~ 0.5
- Protein data:
  - BACC > 0.94 (mixture models)
  - MSA composition highly informative

Discussion

- Remarkable Classification Accuracy ⇒ low simulation realism
- Representative empirical data + as-good-as-possible simulations
- Two distinct approaches = two distinct sets of characteristics
  - GBTs: hand-crafted, MSA global features
  - CNNs: site-compositions
- Important features:
  - Site-composition
  - Uniformity of evolution across sites

Conclusion

- What now?
  - Better models
  - New (model-free?) simulators
- Classification approach as testing framework for simulation realism
  - High accuracy ⇒ low realism
  - Low accuracy ⇏ high realism
- Models ≠ real-world, is that surprising?
  - No, but the degree of unrealism is!