Integrating heterogeneous predictive models using Reinforcement Learning

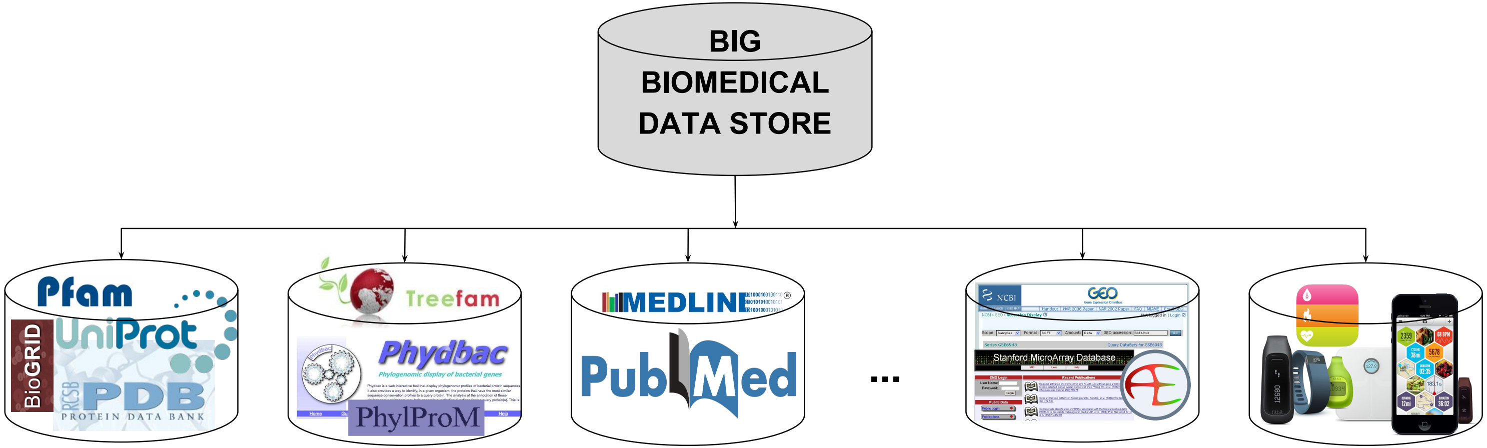
Ana Stanescu[1] and Gaurav Pandey[2]

[1]Department of Computer Science, University of West Georgia, Carrollton, GA, USA

[2]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, USA

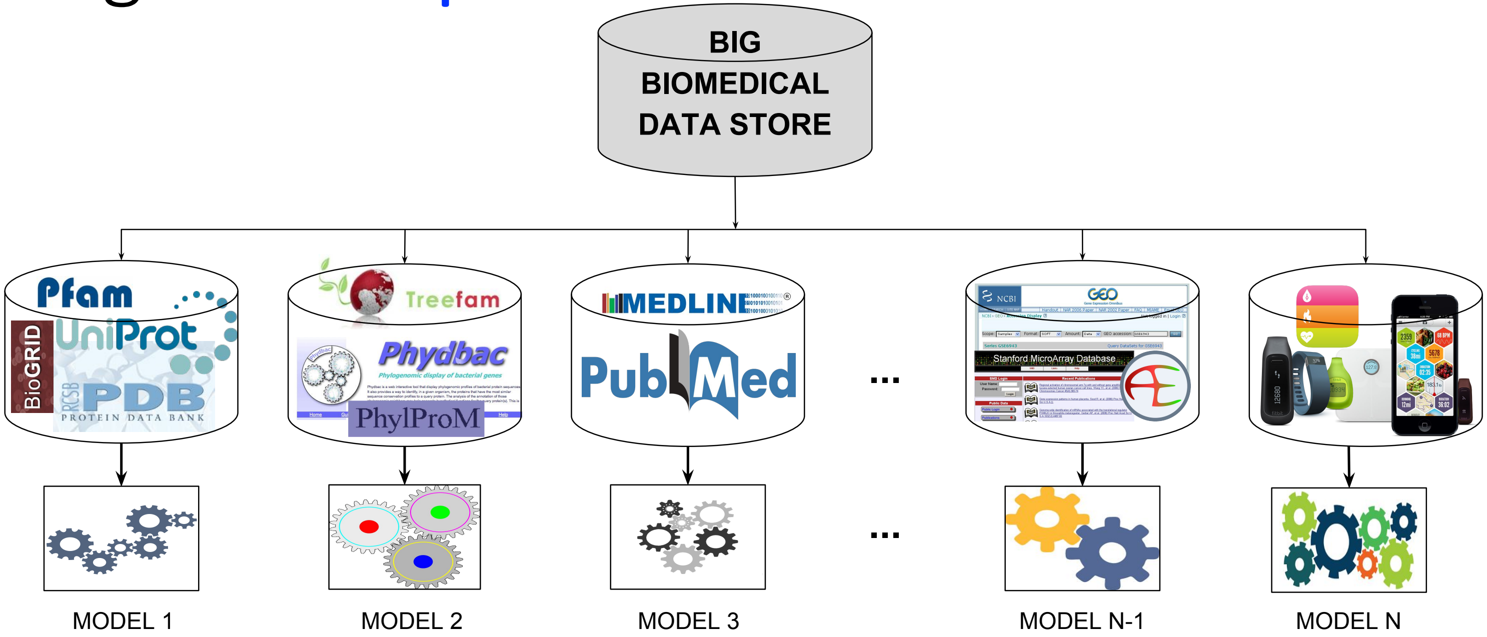
*** Partially supported by the NIH BD2K Program**

Biomedical data are abundant



BIG
BIOMEDICAL
DATA STORE

Systems biology and machine learning can generate predictive models from data



BIG BIOMEDICAL DATA STORE

...

...

MODEL 1

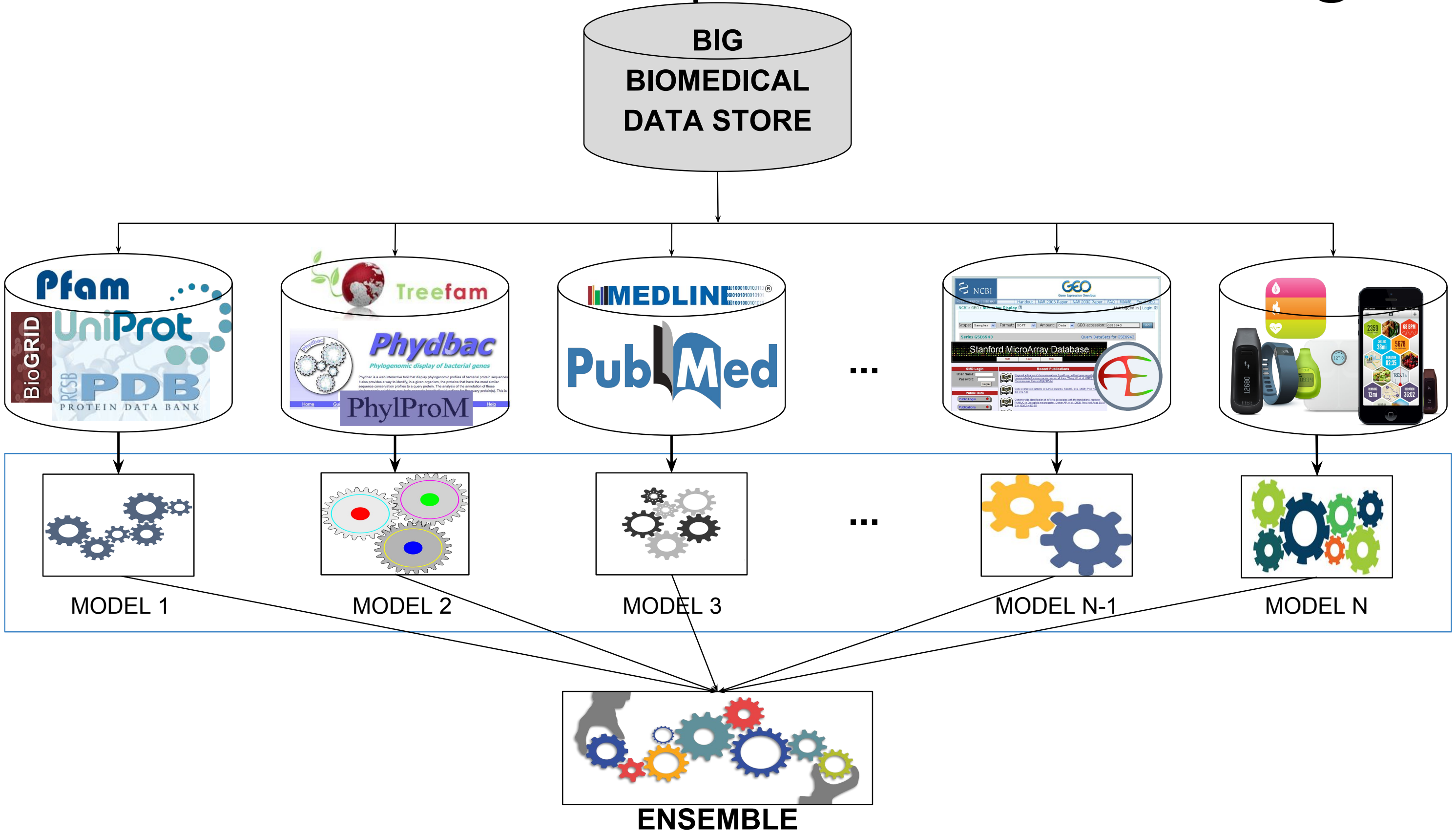
MODEL 2

MODEL 3

MODEL N-1

MODEL N

Heterogeneous ensembles can enhance effectiveness of predictive modeling



BIG BIOMEDICAL DATA STORE

MODEL 1

MODEL 2

MODEL 3

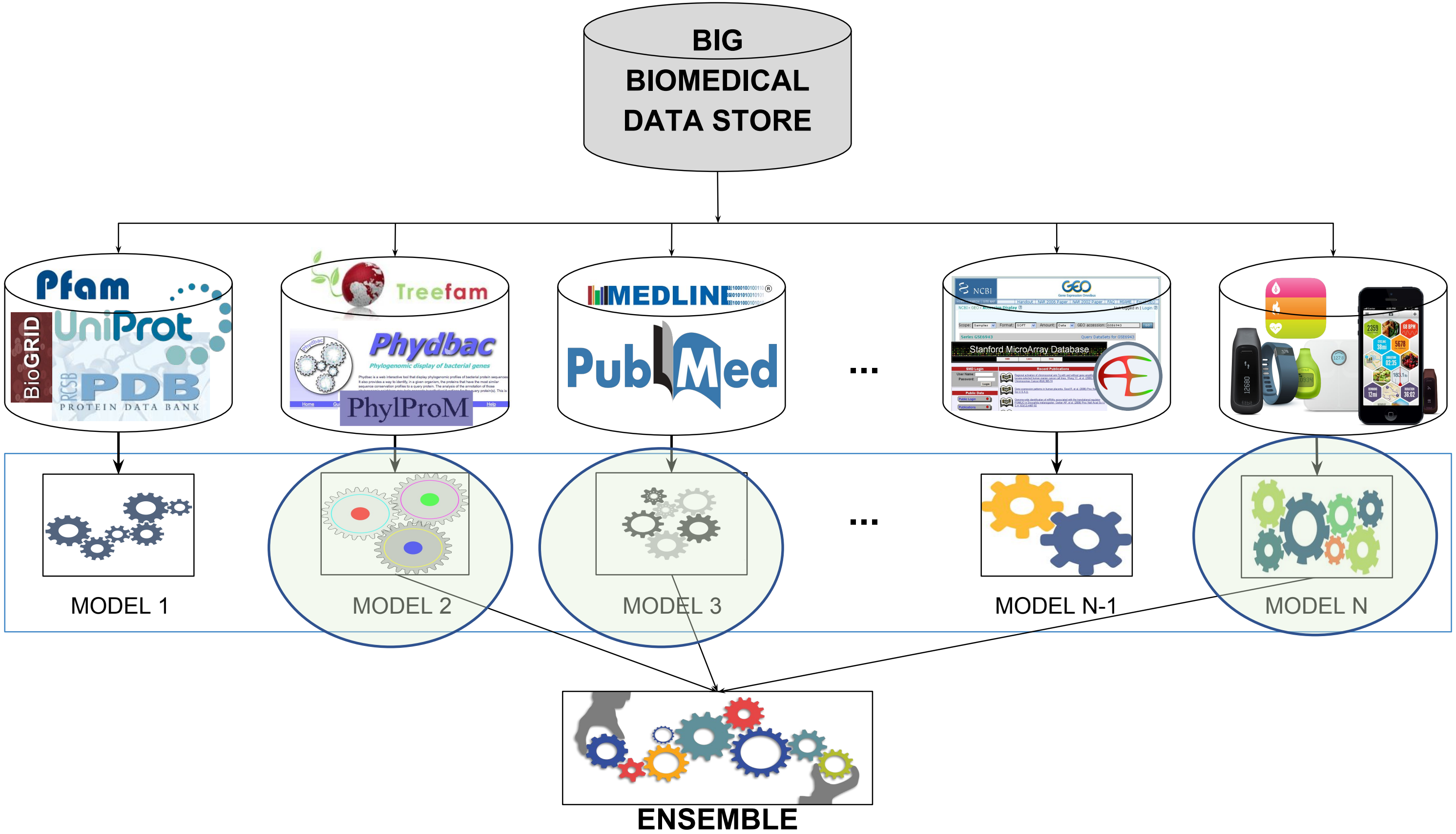
...

MODEL N-1

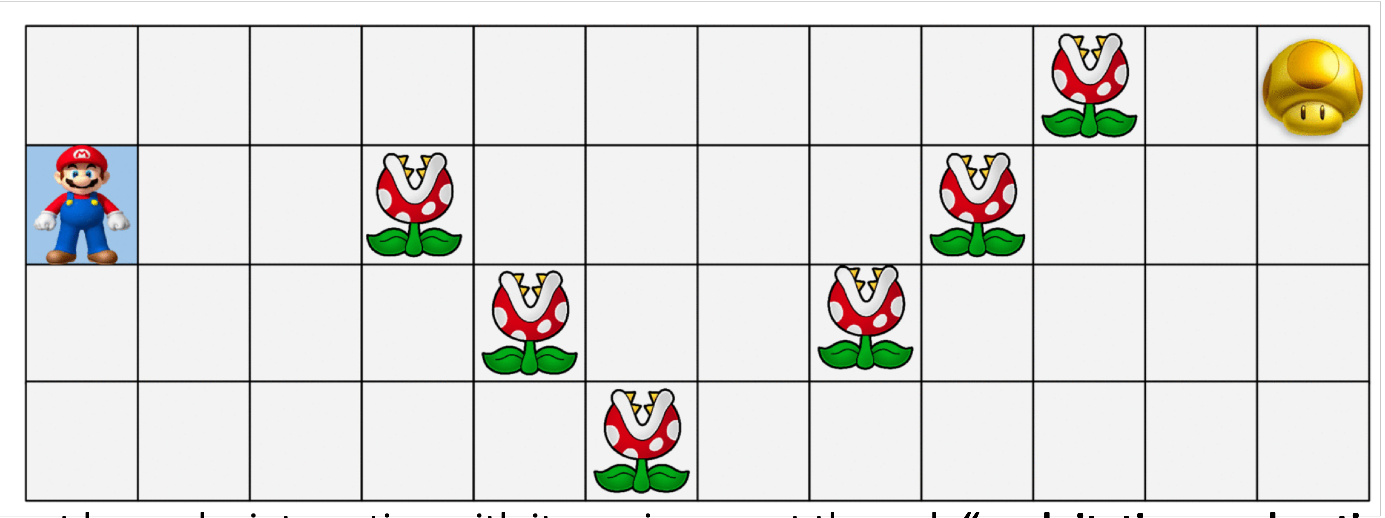
MODEL N

ENSEMBLE

Selecting a parsimonious set of models into an ensemble can further advance predictive performance and interpretability

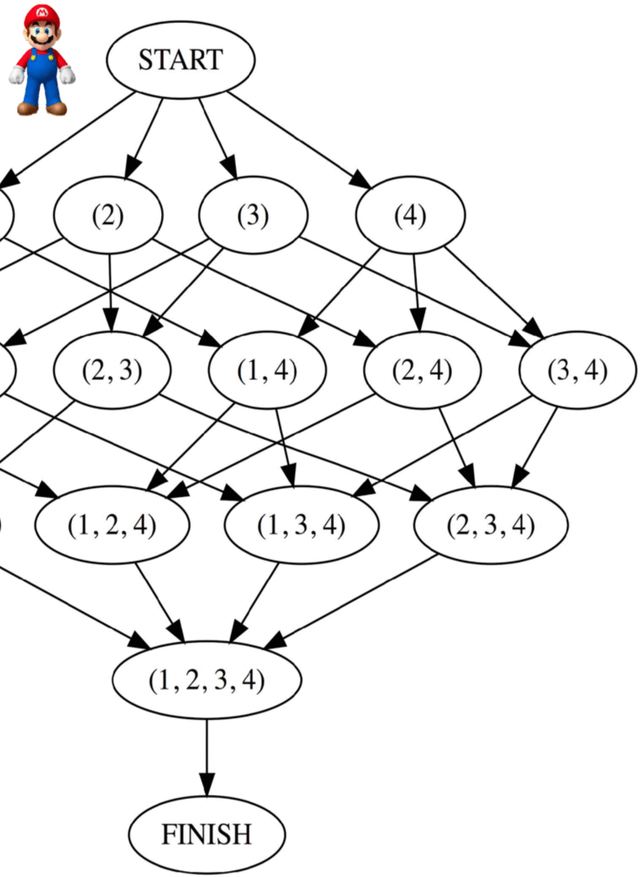


Reinforcement Learning: searching a large structured **environment** with **rewards** to find an optimal path (behavior) to reach the goal



An agent learns by interacting with its environment through **“exploitation-exploration”**.

Ensemble selection using Reinforcement Learning



Reward functions can be formulated in terms of ensemble performance and/or diversity

Fine balance between ensemble performance and ensemble diversity

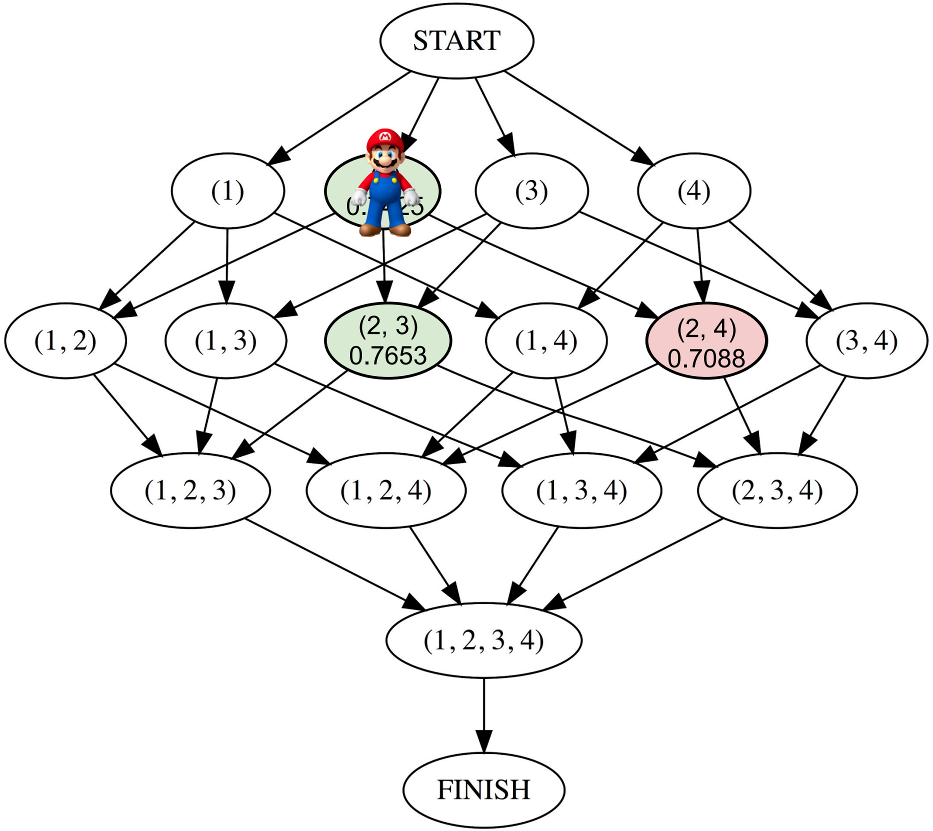
We have designed several search strategies focused on:

- **performance**

  (Stanescu and Pandey, PSB 2017)

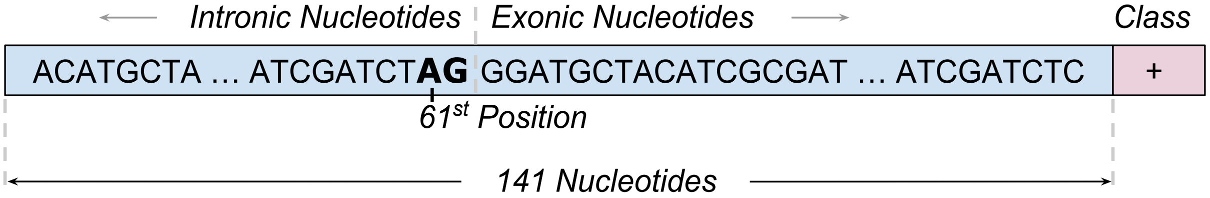
- **diversity**

  (Stanescu and Pandey, arXiv 2018)



Target problem and evaluation methodology

- Predict splice sites in various organisms based on nucleotide positional match representation using several public datasets.



Intronic Nucleotides

Exonic Nucleotides

Class

ACATGCTA … ATCGATCT**AG** GGATGCTACATCGCGAT … ATCGATCTC

+

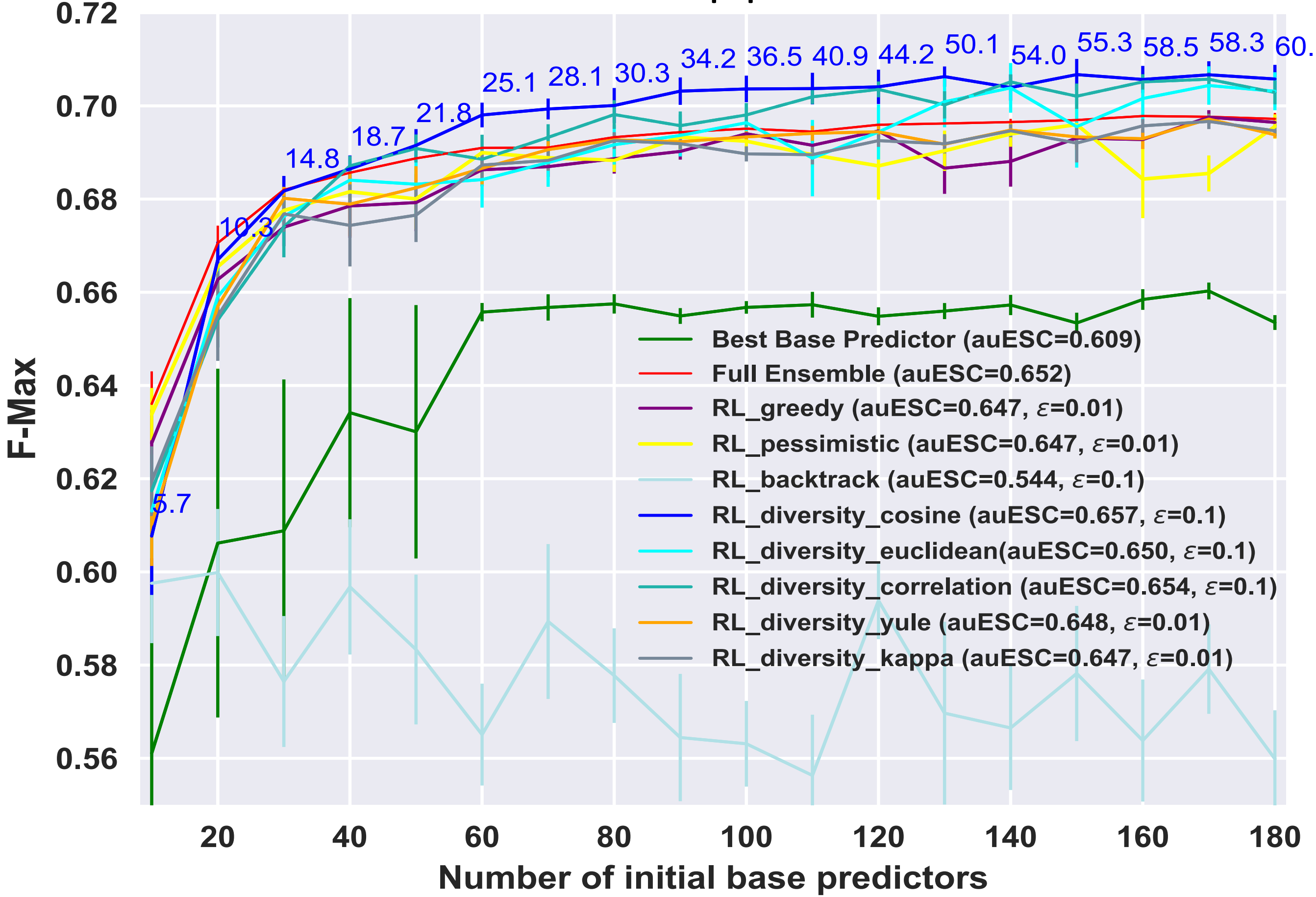
61st Position

141 Nucleotides

Problem	*C. elegans*	*D. melanogaster*	*P. pacificus*	*C. remanei*	*A. thaliana*
#Features	141	141	141	141	141
#Positives	1,598	997	1,596	1,600	1,600
#Negatives	158,150	99,003	156,326	157,542	158,377
Total	159,748	100,000	157,922	159,142	159,977

- 10 bagged versions of 18 different classifiers: 180 base classifiers in a 5-fold cross-validation setup.

Performance of ensembles selected using RL and other approaches



Conclusions

- Reinforcement learning-driven ensembles are **competitive in predictive performance to** larger ensembles consisting of all base predictors, while being substantially smaller, *i.e.*, **more parsimonious.** (Stanescu and Pandey, PSB 2017)

- Ensemble diversity, measured appropriately, can build **even more accurate and parsimonious ensembles**. (Stanescu and Pandey, arXiv 2018)

- **Implementation available**: **https://github.com/GauravPandeyLab/LENS**

- Future Work
  - Test the RL ensemble framework on larger datasets, including non-biomedical ones.
  - Develop more efficient (parallel) implementations of the framework.

Acknowledgements

- NIH Grant R01-GM114434
- Pandey Lab at Mount Sinai
- Minerva supercomputing team at Mount Sinai
- Computer Science Department at UWG

Thank you!