# Active Learning for microRNA Prediction

Mohsen Sheikh Hassani & Dr. James R Green

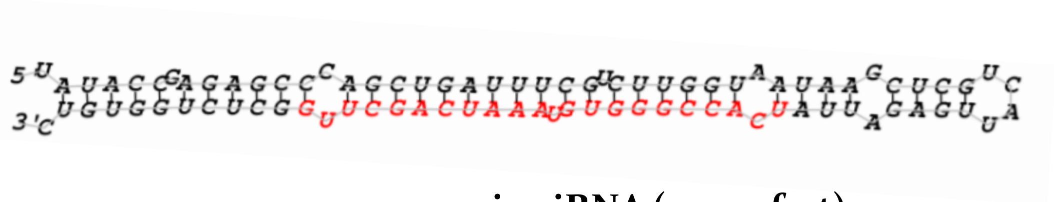
Carleton University

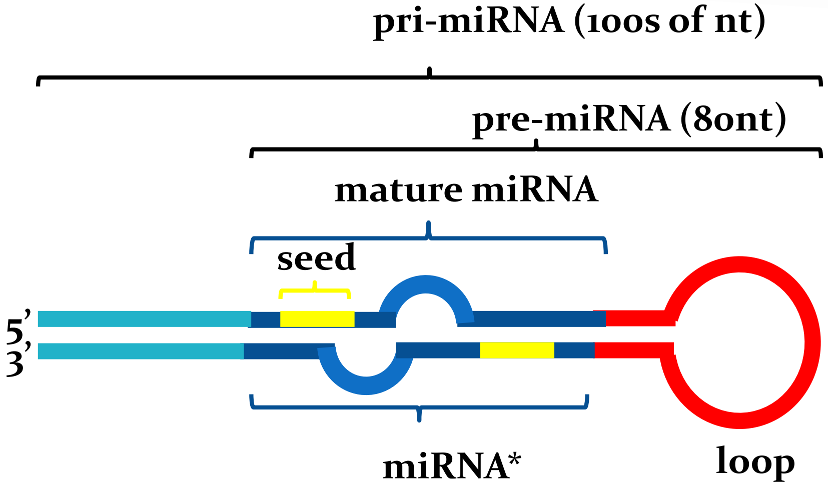
BIBM 2018, Madrid

December 4th, 2018

# MicroRNA (miRNA)

- Short non-coding RNAs
- Typically 18-25 nucleotides
- First miRNA discovered in 1993 (roundworms)
- Next discovery was in 2000
- Today, thousands of known miRNA



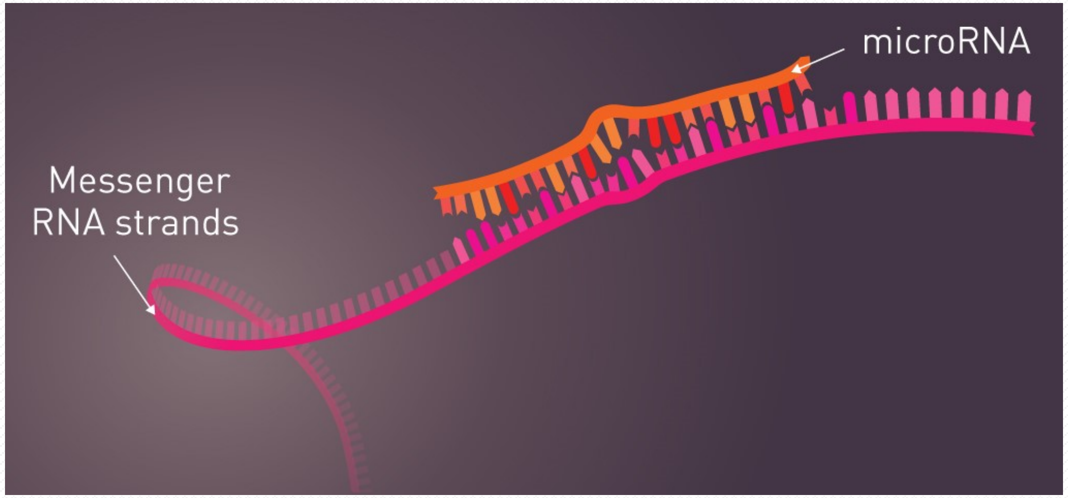


# Why are miRNA important?

- Through gain- and loss-of-function experiments, evidence shows miRNA regulate the expression of proteins involved in:
  - biological development
  - cell differentiation
  - cell cycle control
  - stress response
  - Related to diseases: cancer, neurological disorders, heart disease

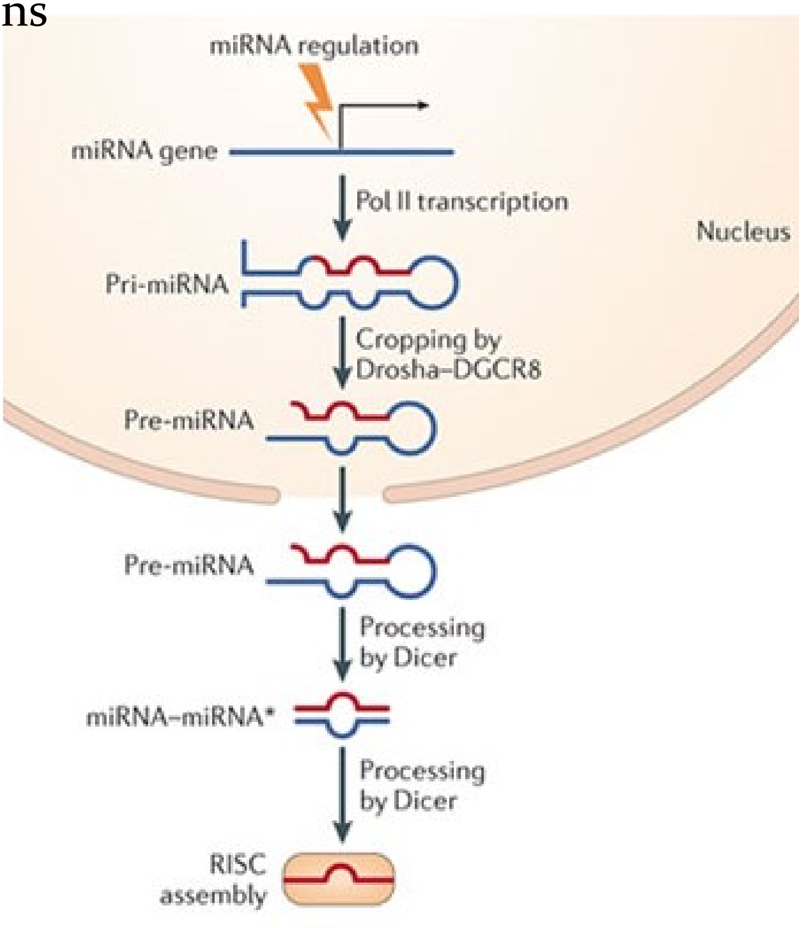
- Predicted to regulate over 60% of transcripts in humans

- May target 60-90% of all mammalian mRNA



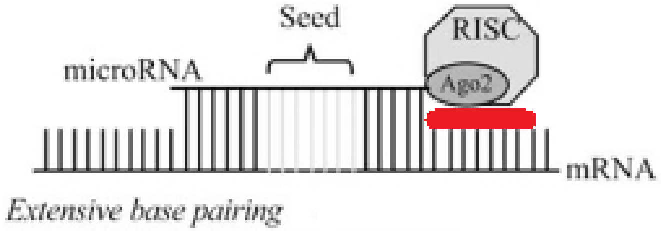
# Biogenesis

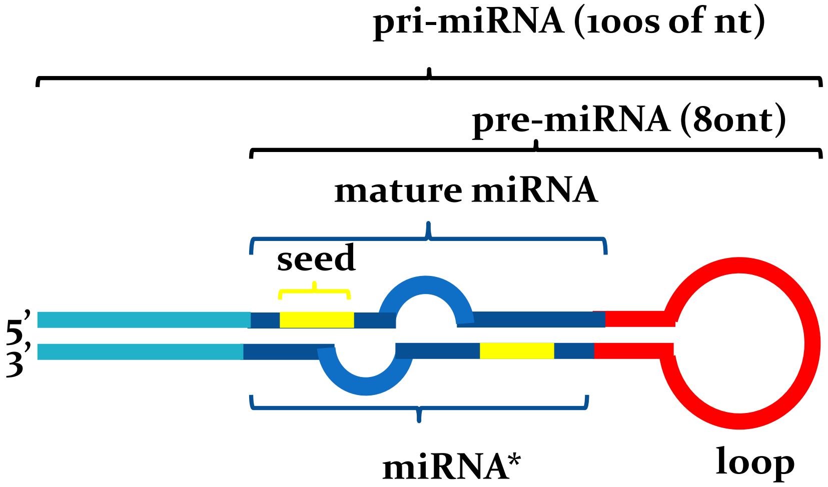
- The biogenesis mechanism plays a key role in miRNA identification
- Either transcribed regions of RNA or introns ( pri-miRNA) fold into hairpins
- Cleaved by enzymes called Drosha in nucleus to ~80 ntds (pre-miRNA)
- Exported to cytoplasm (via Exportin-5 and RanGTP)
- Processed by Dicer ( loop cut off) to ~20 bp
- Two strands of mature miRNA:
  - One strand: Incorporated into miRNA-induced silencing complex (miRISC)
  - Other: Released and degraded



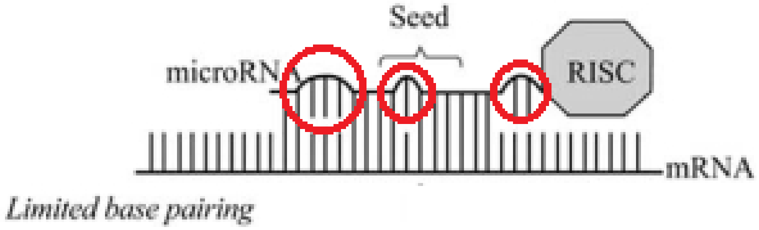
# Gene regulation

- Exact means of miRNA silencing remains unclear.
- Evidence supports two distinct mechanisms:
  - mRNA degradation : miRNA bind to mRNA and promote degradation





  - translation inhibition : miRNA bind to mRNA and prevent translation



# miRNA identification

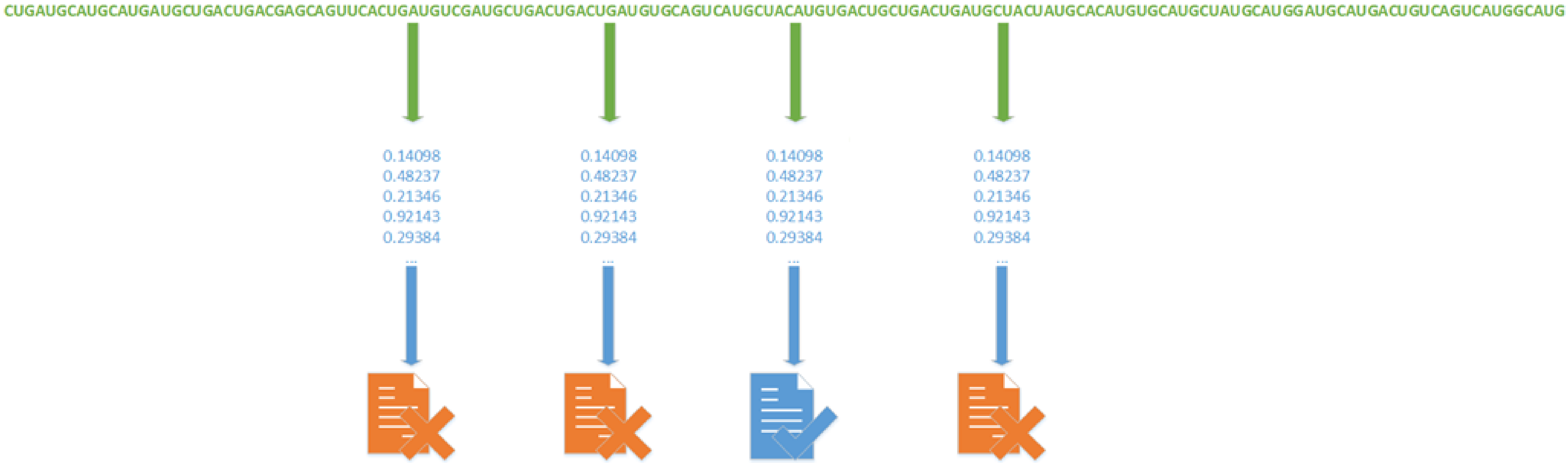
- Requires interdisciplinary strategies; integration of experimental approaches with computational methods

- Computational methods are used to predict, experimental methods are used to validate

- Broadly categorized as either de novo miRNA prediction ( sequence based) or NGS-based (expression-based)

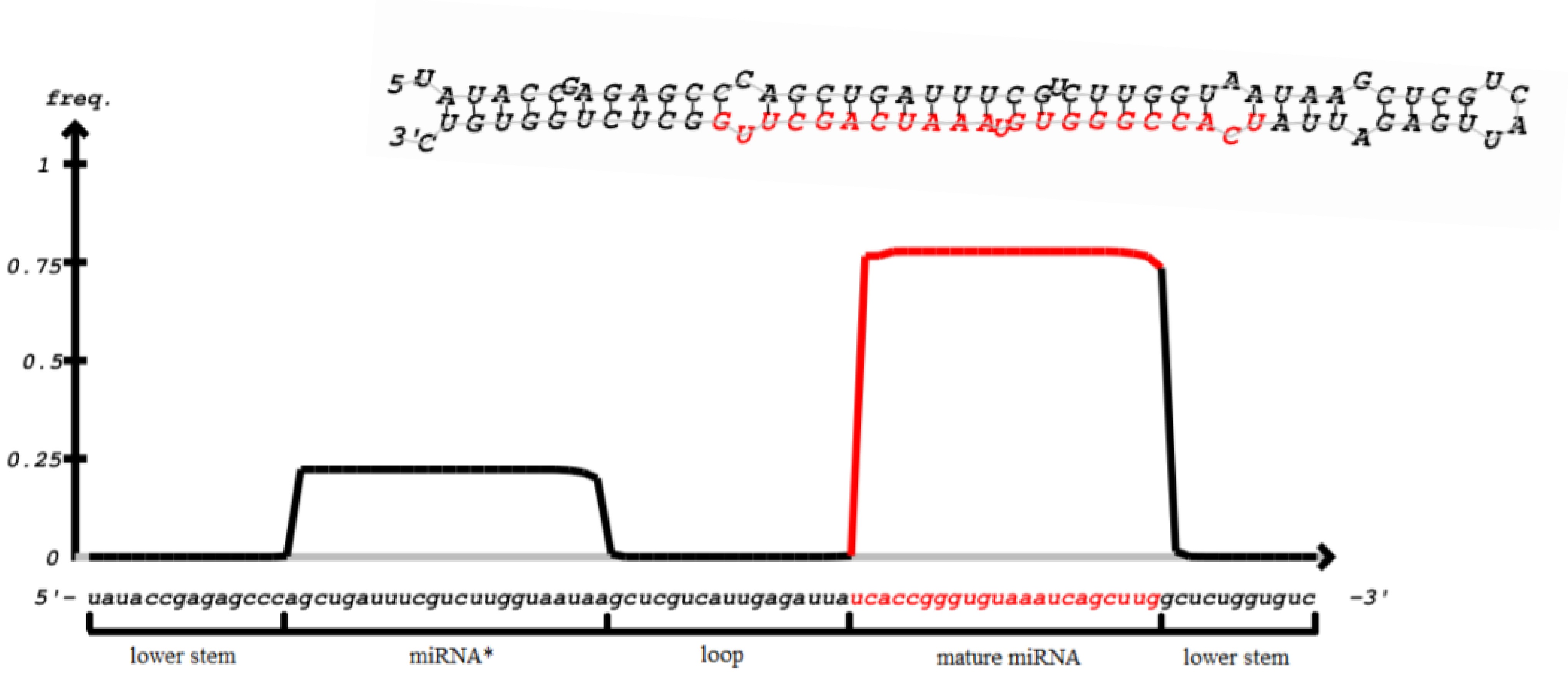
# Computational miRNA prediction

- *De novo* : sequences extracted from genomic data set are classified based on sequence properties
  - Example: look at windows of triplet nts (also single/dinucleotides), how often specific combinations appear



# Computational miRNA prediction

- NGS : Predictions made based on patterns of read depth
  - Example: statistics of the read positions and frequencies of the reads
  - Mature sequences are more abundant in the cell → sequenced more frequently



# Motivation

- miRNA are critical to our understanding of biological processes
  - Identifying greater numbers = better understanding
  - Inter-disciplinary, identification of miRNA remains a difficult task

- Abundance of unlabeled data, scarcity of labeled examples for many species
  - New NGS methods provide large unlabeled data sets

- Existing methods of miRNA prediction require lots of known samples (supervised)

- We wish to extract the most information from limited labelled and available unlabeled data

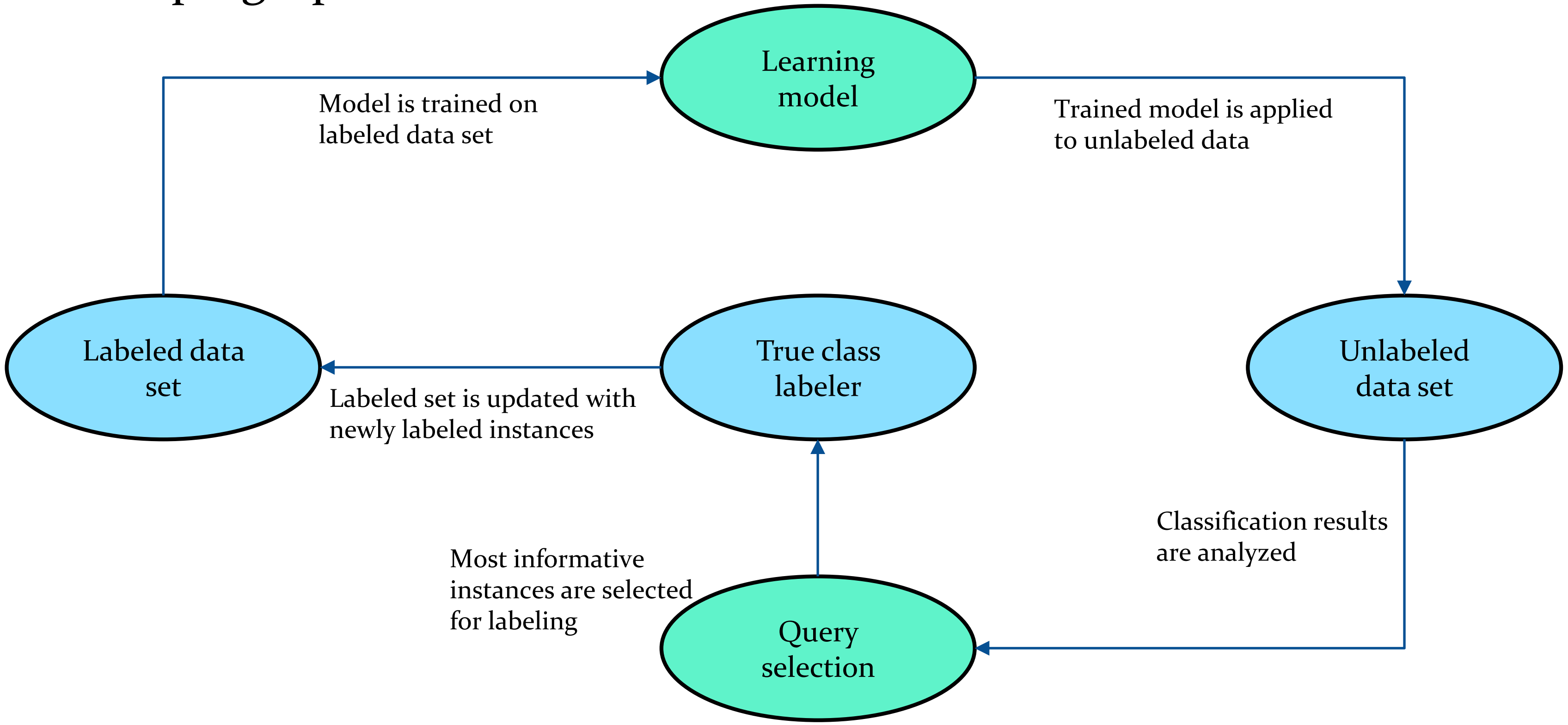
# Problem Statement

- Explore the application of semi-supervised learning (active learning) to miRNA prediction in order to leverage both labelled and unlabelled data.

- Expected Benefits:
  - Require smaller labelled training sets
  - Applicable to more species
  - More value from wet-lab validation experiments

# Active Learning

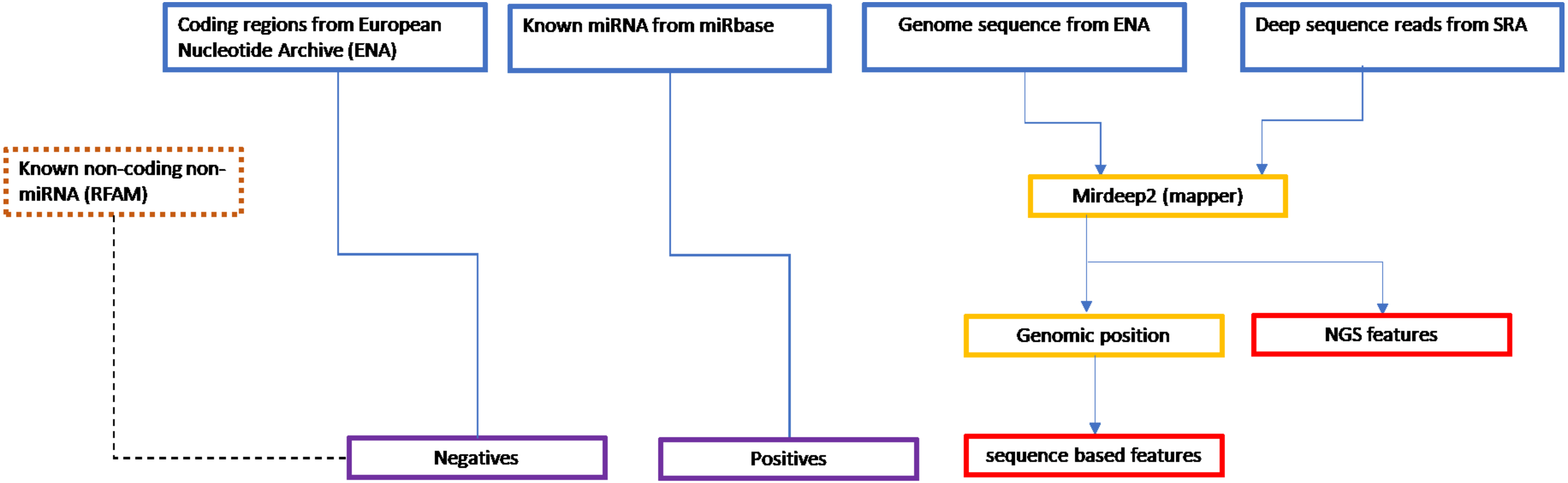
- A semi-supervised machine learning approach
- Interactively query the user
- Suitable when labeling data is expensive
- Minimizes the overall cost of developing a predictor


Learning model
Model is trained on labeled data set
Trained model is applied to unlabeled data
Labeled data set
True class labeler
Unlabeled data set
Labeled set is updated with newly labeled instances
Classification results are analyzed
Most informative instances are selected for labeling
Query selection

# Data Set Creation

- NGS expression data
- Known miRNA
- Known functional non-coding RNA

- Genomic data
- Known coding regions



# Training Data Preparation

- Candidate pre-miRNA that map to known miRNA from miRbase → True positive

- Candidates not identified as miRNA are aligned to coding region data

- Candidates aligning with at most two mismatches are selected as negative samples

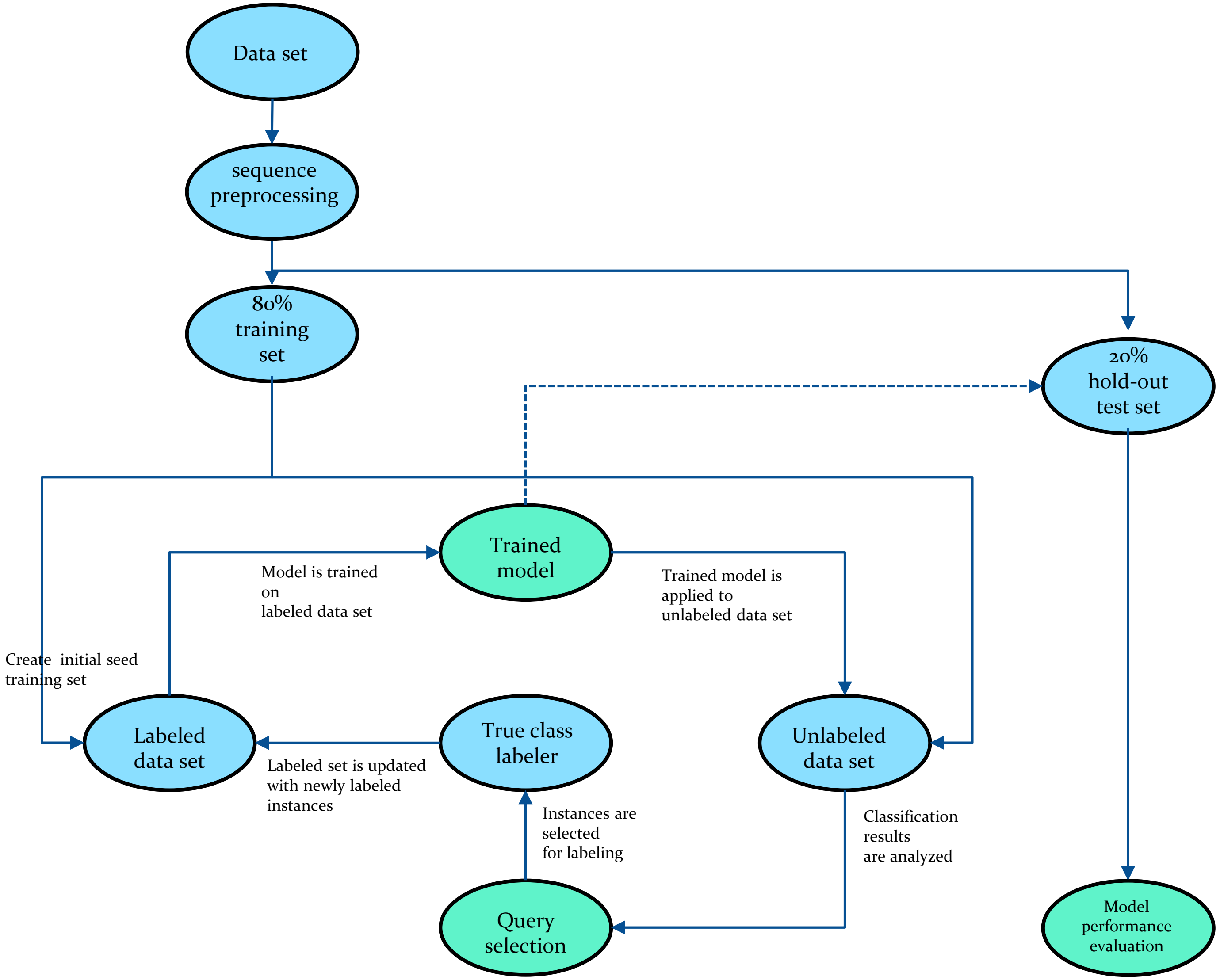
  + known non-coding RNA

| Data set | # of positive samples | # of negative samples |
|---|---|---|
| hsa (human) | 509 | 842 |
| mmu (mouse) | 367 | 844 |
| dme (fruit-fly) | 110 | 97 |
| bta (cow) | 332 | 650 |
| gga (chicken) | 193 | 104 |
| eca (horse) | 364 | 224 |

# Active Learning Pipeline

- Test/train data split (20%-80%)
- Feature set selection  (13-6)
- Initial training set size (10 samples)
- Classifier selection (RF)
- Stopping criterion (11 iterations)

- Query strategy
  - How to spend validation budget?
  - Certainty-based
  - Uncertainty-based

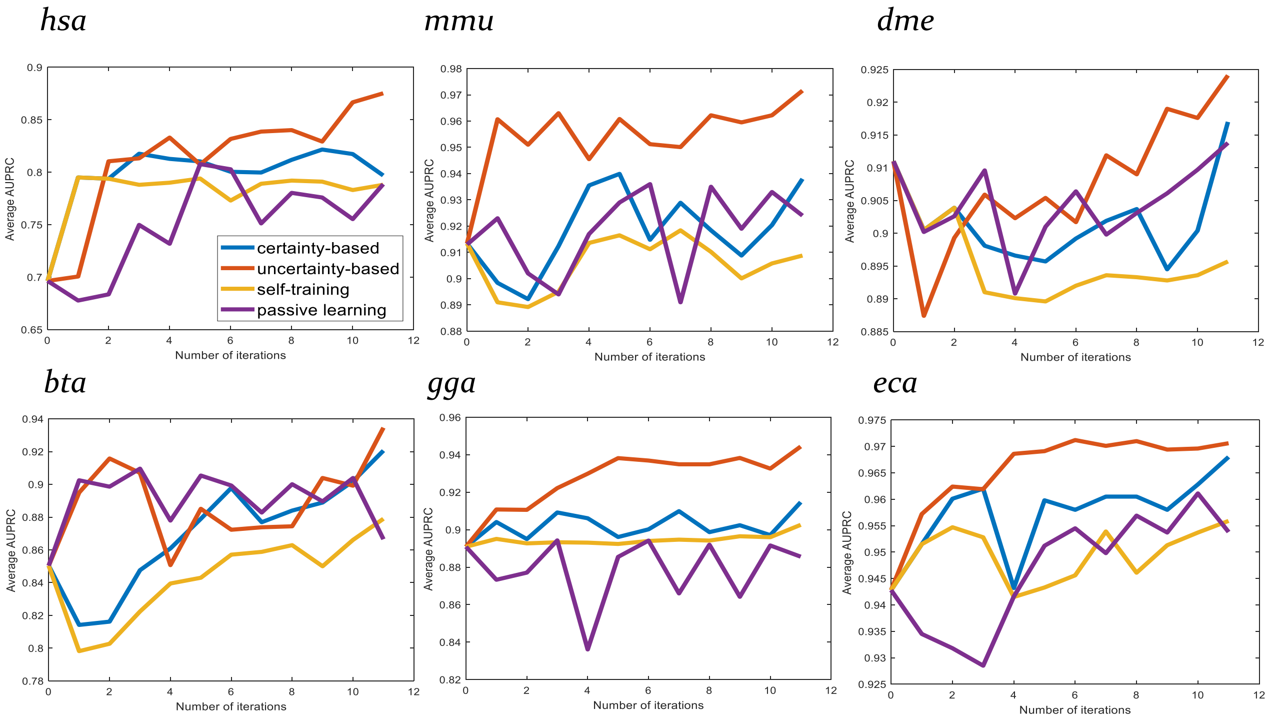

Data set
sequence preprocessing
80% training set
20% hold-out test set
Trained model
Model is trained on labeled data set
Trained model is applied to unlabeled data set
Create initial seed training set
Labeled data set
True class labeler
Unlabeled data set
Labeled set is updated with newly labeled instances
Instances are selected for labeling
Classification results are analyzed
Query selection
Model performance evaluation

# Results

- Active Learning
  - Certainty-based active learning
  - Uncertainty-based active learning

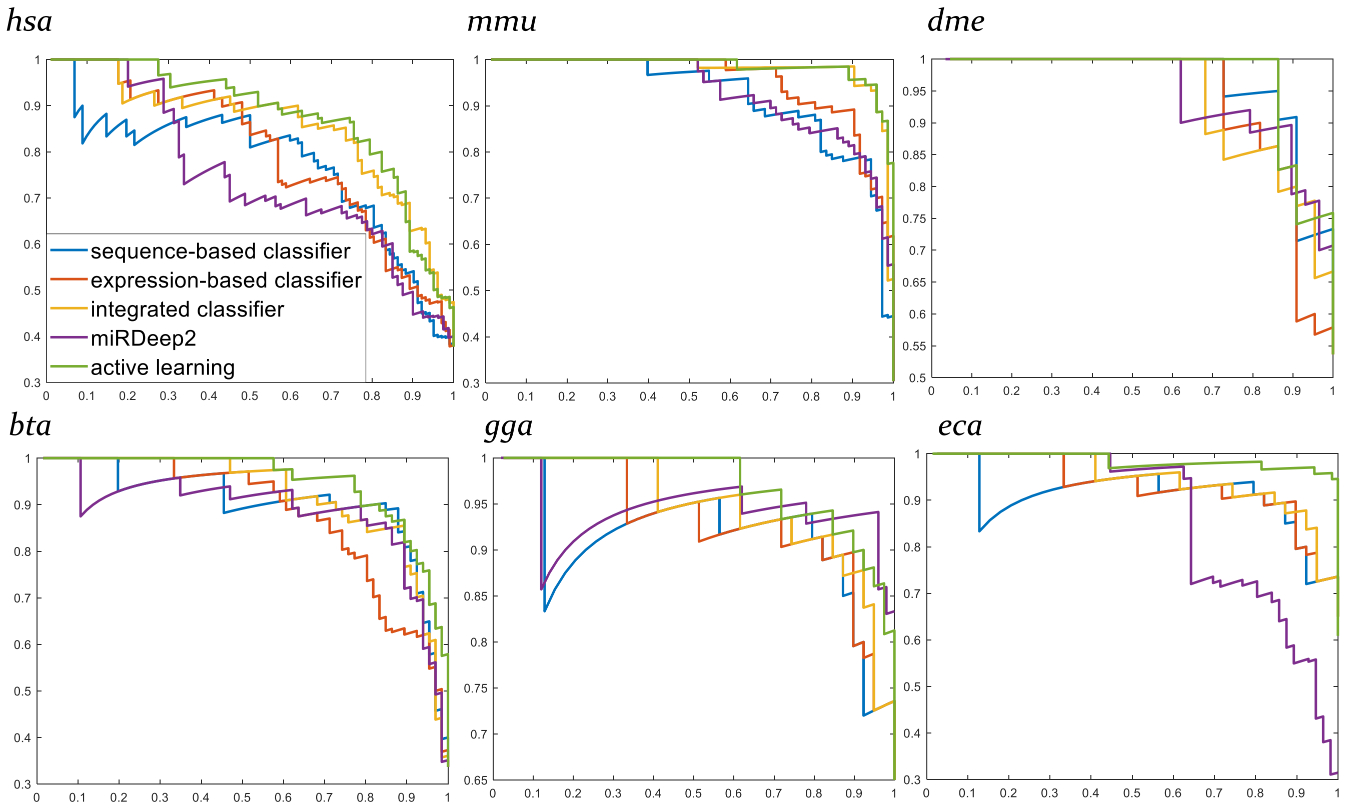
- Baseline methods
  - Self-training
  - Passive learning

| Data set | Self-training average AUPRC | Passive learning average AUPRC | Certainty based average AUPRC | Uncertainty based average AUPRC |
|---|---|---|---|---|
| hsa | 0.788 (+13.1%) | 0.789(+13.2%) | 0.797 (+14.4%) | 0.875 (+25.7%) |
| mmu | 0.909 (-0.50%) | 0.924(+1.16%) | 0.938 (+2.69%) | 0.972 (+6.37%) |
| dme | 0.896 (-1.68%) | 0.914(+0.30%) | 0.917 (+0.66%) | 0.924 (+1.44%) |
| bta | 0.879 (+3.36%) | 0.867(+1.89%) | 0.921 (+8.25%) | 0.935 (+9.90%) |
| gga | 0.903 (+1.31%) | 0.886(-0.60%) | 0.915 (+2.67%) | 0.944 (+6.01%) |
| eca | 0.956 (+1.39%) | 0.954(+1.17%) | 0.968 (+2.67%) | 0.971 (+2.95%) |
| Avg. | + 2.83% | +2.86% | +5.23% | +8.72% |



# Results - continued

| Data set | Sequence-based average AUPRC | Expression-based average AUPRC | Integrated (miPIE) average AUPRC | miRDeep2 average AUPRC | Active learning average AUPRC |
|---|---|---|---|---|---|
| hsa | 0.763 (±0.02) | 0.789 (±0.01) | 0.844(±0.01) | 0.736 | 0.875(±0.01) |
| mmu | 0.907 (±0.01) | 0.939 (±0.01) | 0.966(±0.01) | 0.915 | 0.972(±0.00) |
| dme | 0.918 (±0.01) | 0.893 (±0.01) | 0.894(±0.01) | 0.914 | 0.924(±0.01) |
| bta | 0.890 (±0.02) | 0.865 (±0.02) | 0.905(±0.02) | 0.869 | 0.935(±0.01) |
| gga | 0.886 (±0.02) | 0.906 (±0.01) | 0.919(±0.01) | 0.923 | 0.944(±0.01) |
| eca | 0.886 (±0.01) | 0.906 (±0.01) | 0.919(±0.01) | 0.843 | 0.971(±0.00) |
| Avg. | 0.875 | 0.883 | 0.908 | 0.867 | 0.935 |



In all plots, the y-axis represents precision while the x-axis is recall.

# Conclusions

- Novel active learning approach for the classification of miRNA
- Decreased the number of labeled samples required
- Targeted the problem of limited known data and made use of unlabeled data
- Improved on state-of-the-art performance

# Future Work

- Development of high-quality integrated training data sets
  - Pooling multiple NGS datasets to cover multiple conditions

- Experimental validation of predictions

# Thank You For Your Attention