CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

MUNI

Modeling Small RNA binding rules using Machine Learning

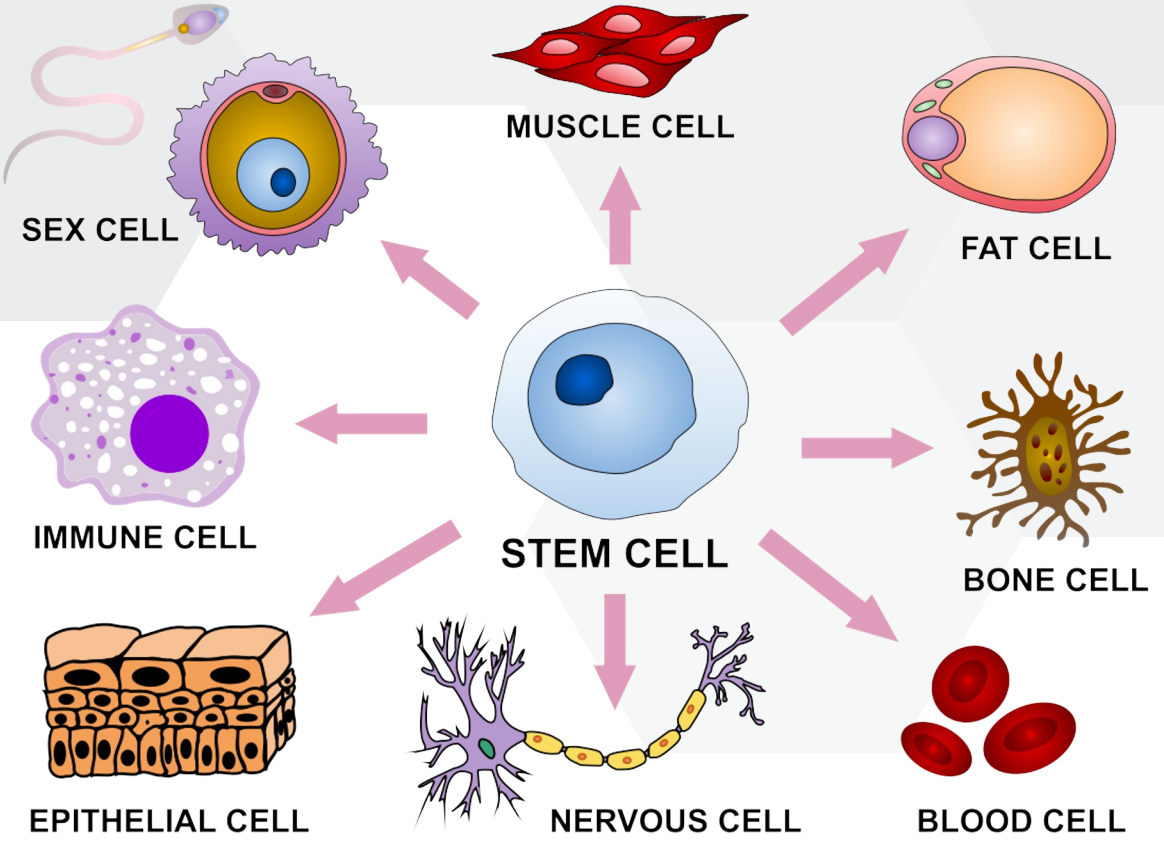
Katarína Grešová

Outline

- Biological background
- Data description
- Current state of the art
- Proof of concept work
- Ideas and plan

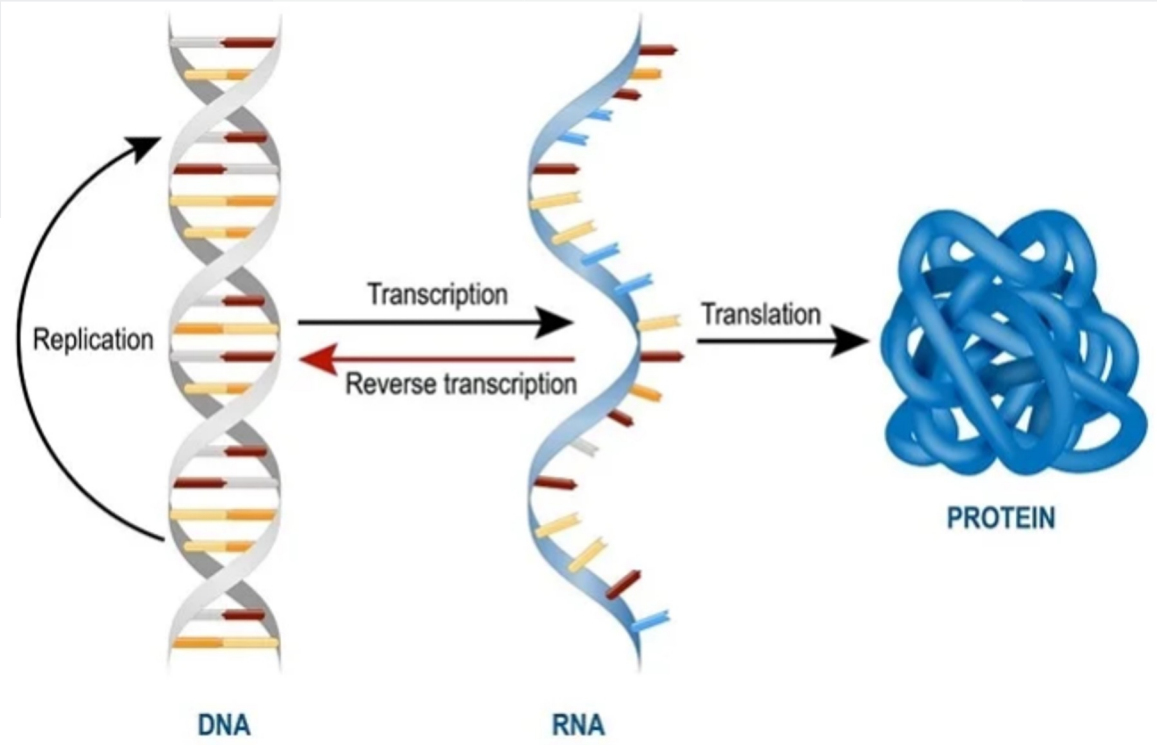
Cells

- Single cell organisms
  - Whole life in one cell
- Multi cell organisms
  - Different types of cells
  - But each cell has exactly the same instructions (DNA) for operating



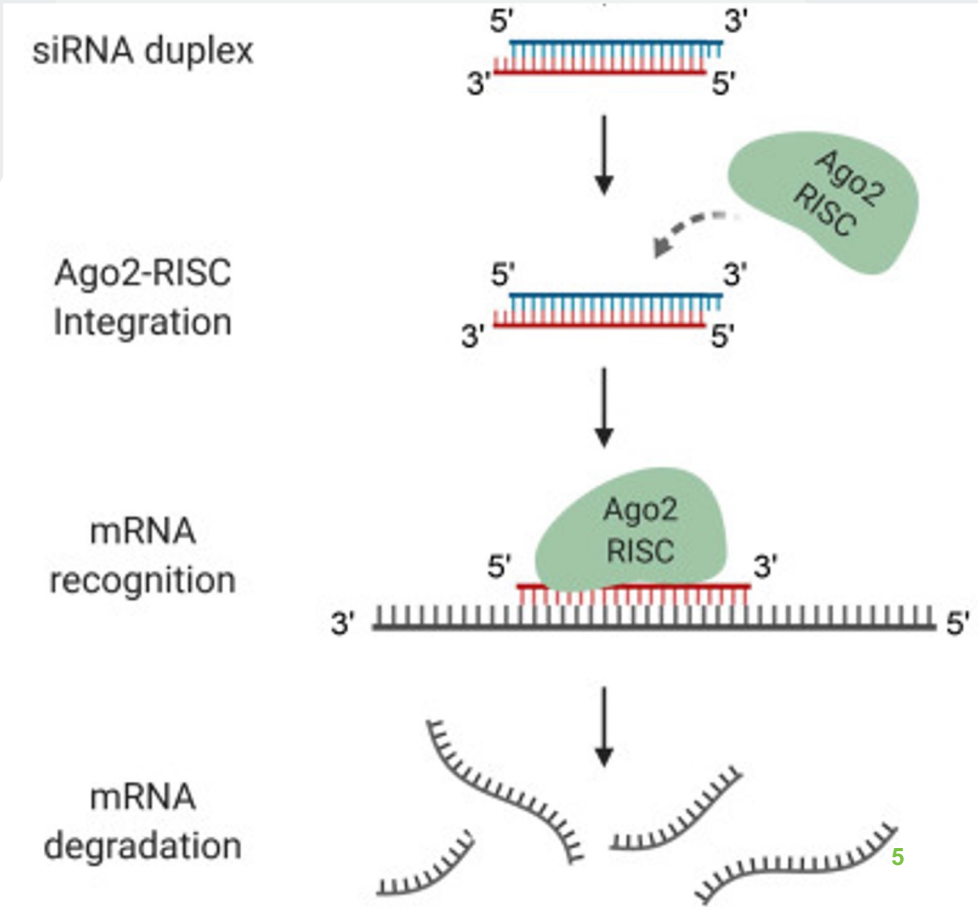
Central Dogma of Molecular Biology

- DNA
  - stores all the instructions for functioning of a cell
- RNA
  - created as a copy of some instruction from DNA
  - can be used to create Protein or it can do some work in cell in a form of RNA
- Protein
  - product of instruction stored in RNA
  - do most of the work in the cell



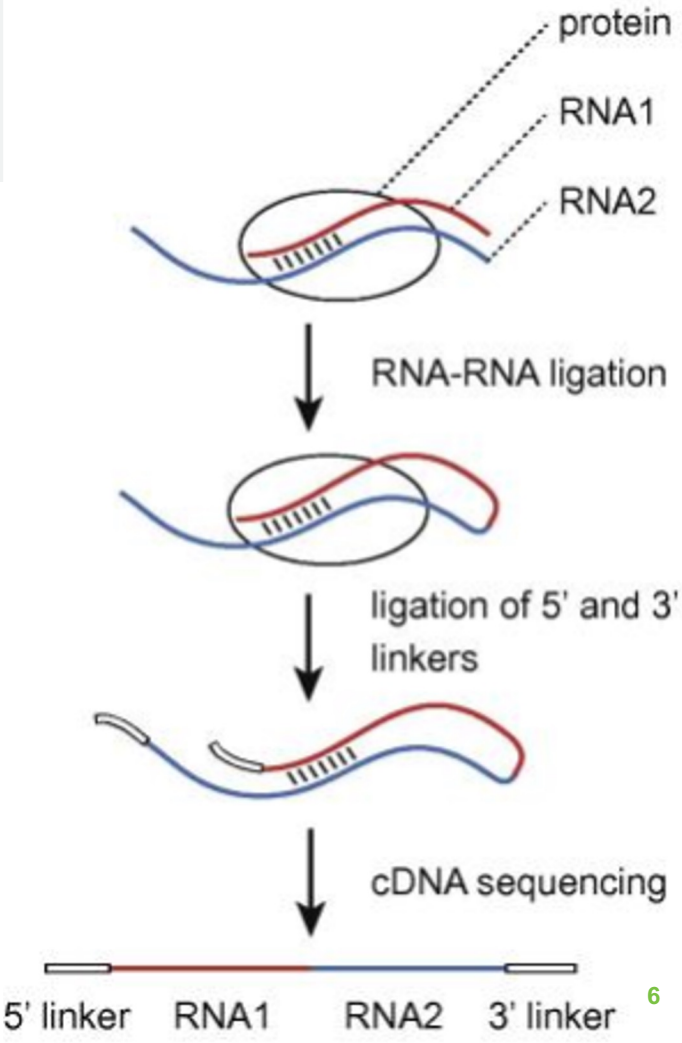
RISC (RNA-induced silencing complex)

- RNA note (called mRNA) transcribed from DNA floats in cell
- Ago protein attaches small RNA onto itself
- Ago uses this small RNA to find specific mRNA
- Identified mRNA is destroyed



Biological experiment

1. Find RISC complex (Ago, small RNA, target RNA)
2. Connect ends of small RNA and target RNA
3. Remove Ago
4. Add specific sequences to ends of connected RNAs
5. Read connected RNAs using PCR



Data

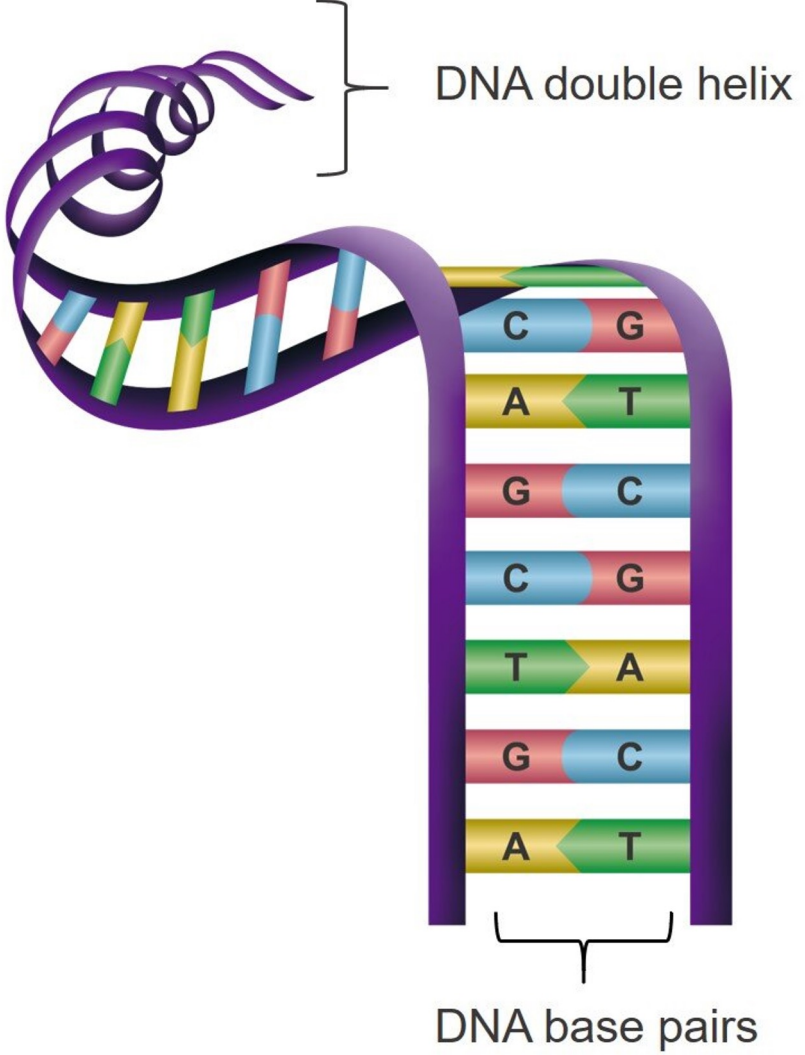
30 785 samples

miRNA	gene	label
AACTGGCCCTCAAAGTCCCG	TGGAGAGCGGGCTTAAGAAGTGGCGGTTCGGCCGGAGGTTCCATCGTATC	1
ATCAGGGCTTGTGGAATGGG	CTCGCTGGCGTTCTCCGGGGTGGTTGGCATTGTGTCCTGGAAGCGGCCAT	0
TGGGGAGCTGAGGCTCTGGG	CTACACCTCAGCCCGGGGCTGCACTGCCACCCTGGGCAACTTCGCCAAGG	0
GTGAGGGCATGCAGGCCTGG	GTAAGGAGCTGGAGTCGCTGGTAGAGAACGAGGGCAGTGAGGTGCTGGCG	0
ATGCACCTGGGCAAGGATTC	GCATATGGGGGCCTTAAGGAATAACAGTGTGCGTGGTGGTGTGCAGGAGA	0
TGCACGGCACTGGGGACACG	TCAGGGTTTCTTGGGGGCTTATGAGTCTCACCGGTCAACCCAGGAGGCCT	0
AACTGGCCCTCAAAGTCCCG	ACCTCTTAATGGGCCAGTGAATAACACTCACTGCTGGCATTTAATGTGCA	1
TGGGTTCCTGGCATGCTGAT	CACCTGCTGCCCCTTCTACCCCAGCTCCACCACCTGCAGTCCCTAAAGAA	0
TCAGTGCATCACAGAACTTT	ACCCGCACAGCAAGCACCTGTACACGGCCGACATGTTCACGCACGGGATC	0
CTGGCCCTCTCTGCCCTTCC	CTGATTGTGGCAGAGGGGCCACTACCCAAGGTCTAGCTAGGCCCAAGACC	1
TGAGGTAGTAGGTTGTATAG	ATGACCCAACCTACCACCCTGTTTTTACATATCCAATTCCAGTAACTCTC	1
TAAAGTGCTTATAGTGCAGG	CAAAAGCATACCTACCTTCCCCTAGAGGTCTGTAACATTGTGGCTGGGCA	1
TGAGAACTGAATTCCATGGG	CCTGGGACCCCCAGGCGTGGAGGACAGTCAAGCCGTGGAGGCCGTGGAGG	0
TGAGGTAGTAGGTTGTATAG	CCCAACCTCAACCTCAACCTCCCAGCACCACACATCATGCCAGGGGTTGG	1
CTGTACAGGCCACTGCCTTG	GAAGGTAAAGAGGGTCATTGGGGTCGAGCTATGCCCAGAGGCTGTGGAGG	0
GTCCCTCTCCAAATGTGTCT	GCTGGCCAGCGGACTTCTGGAGTTAGCCTTTGCTTTTGGAGGACTGTGTG	0
TTAGGGCCCTGGCTCCATCT	ACACAGGAAGAGGAGCCAGGCCCTTGTACCTATGGGATTGGACAGGACTG	1
TAGGTAGTTTCATGTTGTTG	TCCGCCCTCTTTTGCCAGCCCAGCCCCTCCATGCACATTTGGACGCTGTC	0
TAAAGAGCCCTGTGGAGACA	TCCTGAGGCCTGGGGCACCTTTCGTCTGATGAGCCTCTGCATGGAGAGAG	0
GTGGGTACGGCCCAGTGGGG	CATCTTGTCCTCACAGCCCAGAGCATGTTCCAGATCCCAGAGTTTGAGCC	0

Base pairing

A – T

G – C

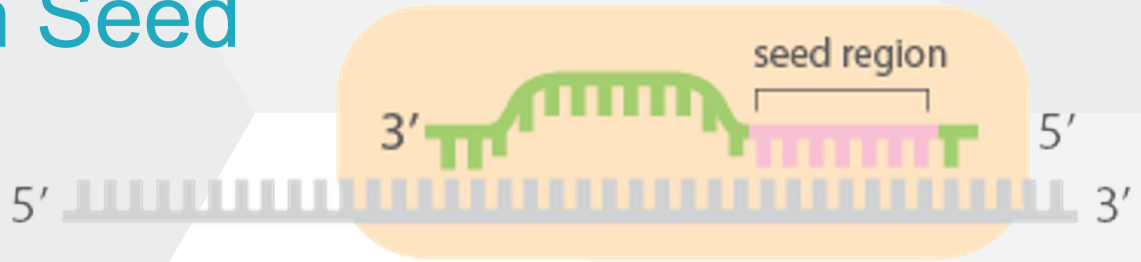


Some of current solutions

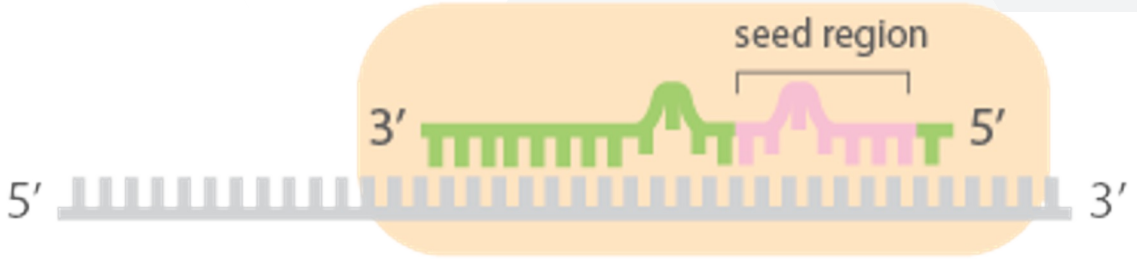
- Based on seed
- Based on base pairing

Solution based on Seed

- Heuristic
- Looks for base pairing in seed region
- Finds only 40% of interaction

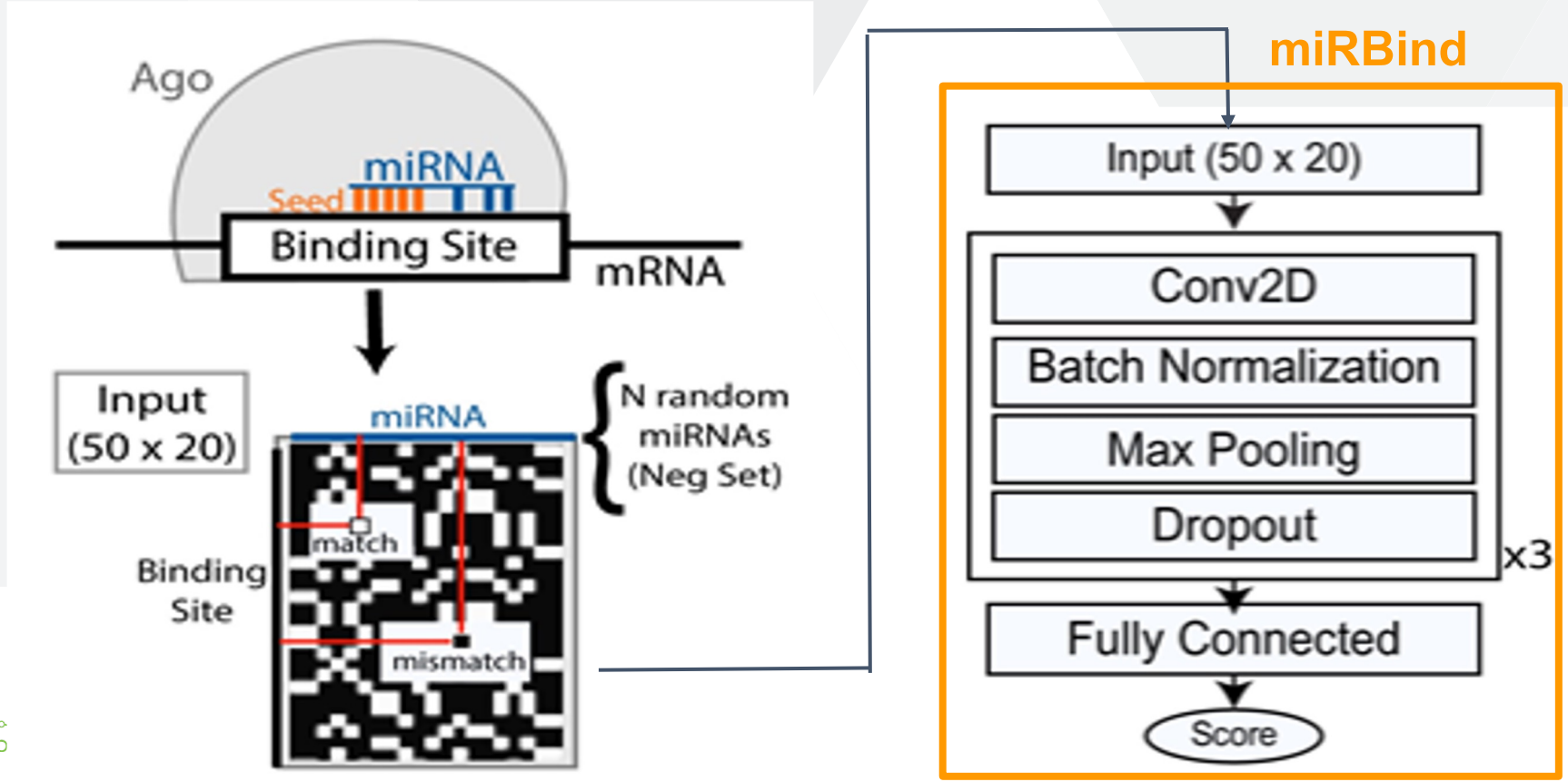


Perfect binding to mRNA



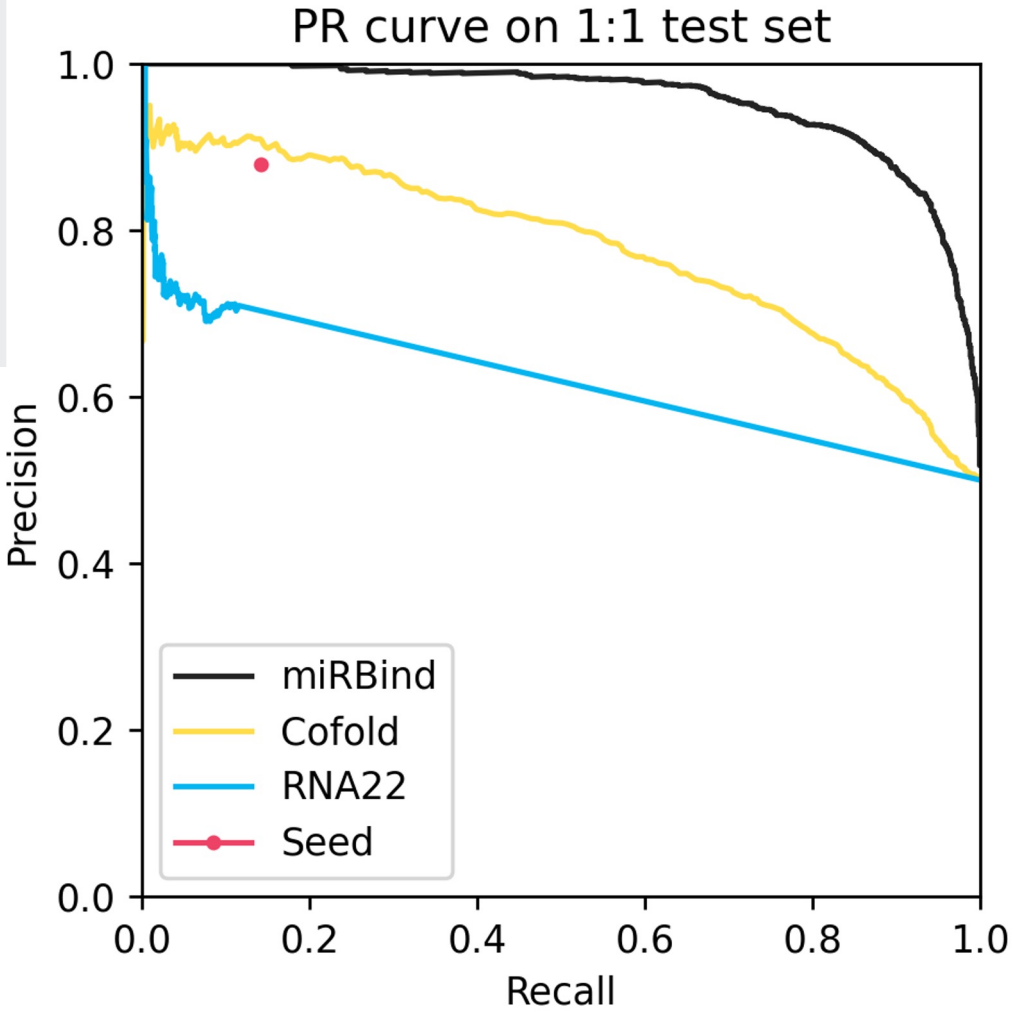
Imperfect binding to mRNA

Solution based on base pairing



Comparison

- **miRBind** - base pairing method (the best results)
- **Cofold** and **RNA22** - not mentioned methods, but also used for prediction
- **Seed** - seed based method, most used



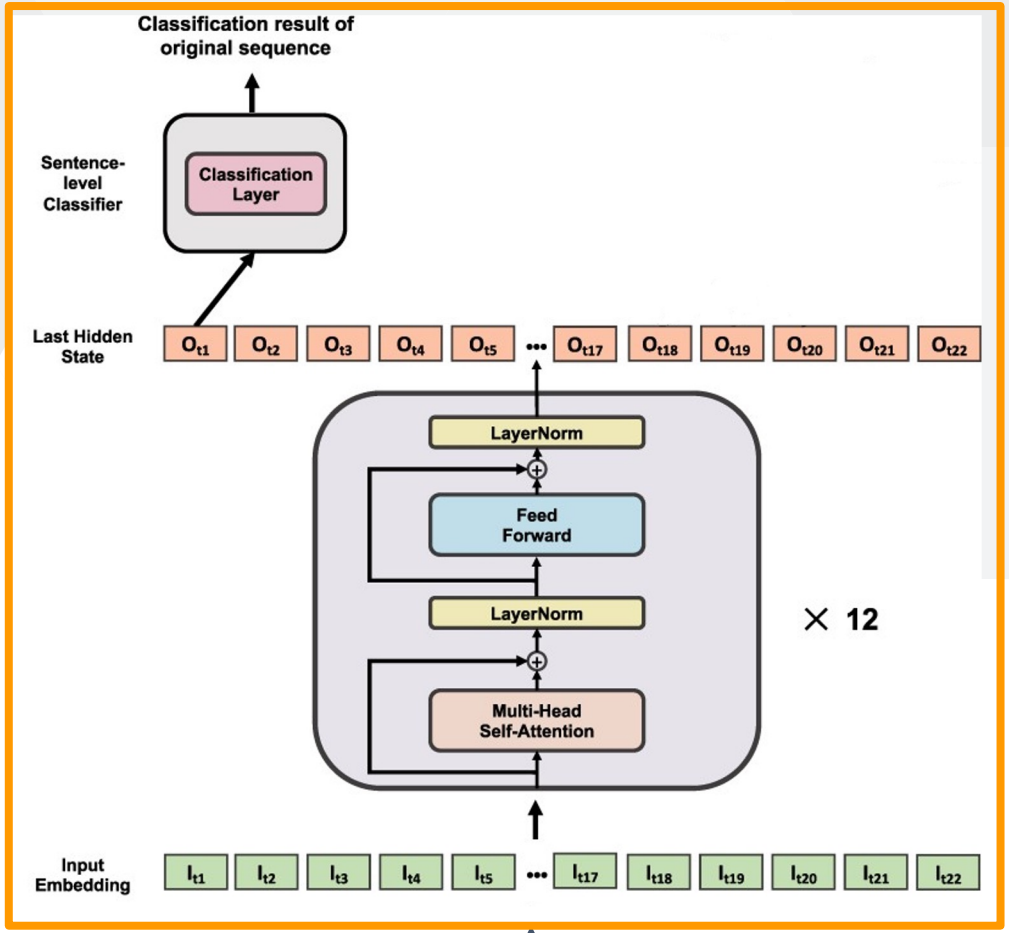
My approach

- Sequence
- Pretrained BERT model - DNABERT

Proof of concept

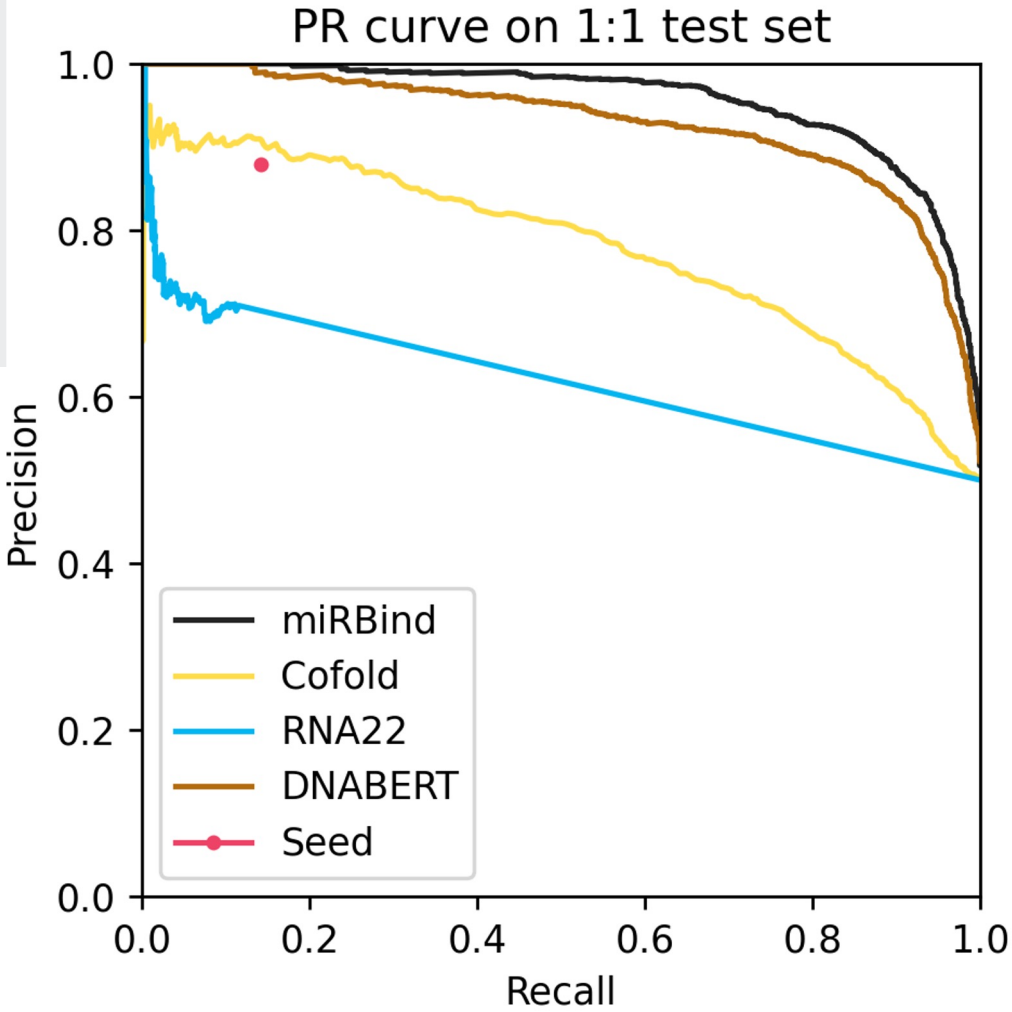
Input
(50+4+20)

ACATTCATTGTTNNNNGCAGGTCTGACTGCAGAGA



First try

- miRBind - base pairing method (the best results)
- Cofold and RNA22 - not mentioned methods, but also used for prediction
- **DNABERT** - sequence based, finetuned DNABERT (mine)
- Seed - seed based method, most used



Experiments

- Batch size 12, 32 and 64
  - No significant change
- Using [SEP] token instead of ‘NNNN’
  - No significant change
- Trained from scratch
  - Worse result

Ideas and Plans

- Do hyperparameter search - how to do it optimally?
- Use RNABERT instead of DNABERT
- Pre-train my own model based only on RNA
- Add training tasks (DNABERT did only [MASK] prediction)
  - Next sequence prediction as in BERT
  - Structure prediction as in RNABERT