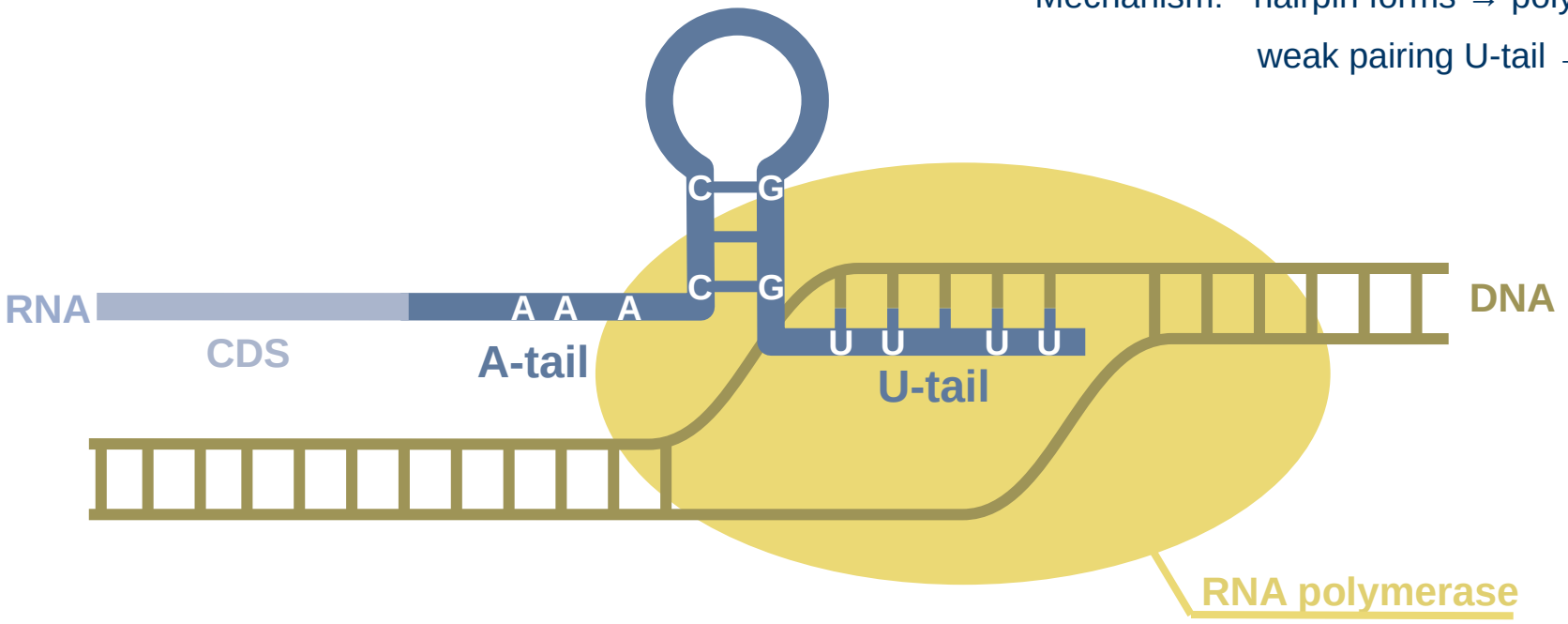
Inverse folding based pre-training for the reliable Identification of intrinsic transcription terminators

Vivian Brandenburg

STUDATA 2021

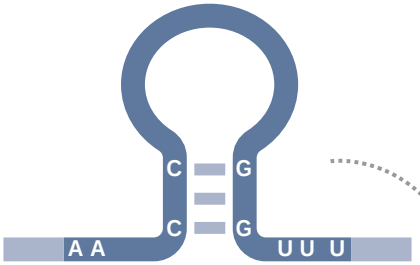
Intrinsic Transcription Terminators

- RNA element on end of nascent transcripts
- Motif: sequence (A-/U-tail) + structure (hairpin)
- Mechanism: hairpin forms → polymerase pauses
  weak pairing U-tail → RNA escapes



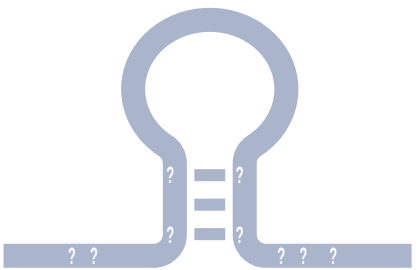
Data: inverse-folding

terminator



+

inverse-folded



- 1175 terminators from *B. subtilis*[1] and *E. coli*[2]
- Used for training of Neural Network
- Limited data availability

- Additional data with same structure but different sequence
- Gained with inverse_fold[3]

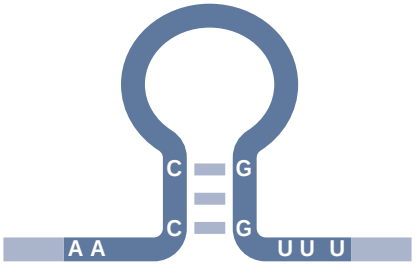
terminator structure	. . **(((** . . . **)))** . .
random sequence	n n n n n n n n n n n n n
inverse-fold sequence	n n **U C C** n n n **G G A** n n

- **Pre-train** with inverse folded data first, then with terminators

1: De Hoon *et al*., PLOS Comp. Biol. 2005 | 2: Chen *et al*., Nature Methods 2013 | 3: RNAlib-2.4.17, Lorenz *et al.,* Algor. Mol. Biol., 2011

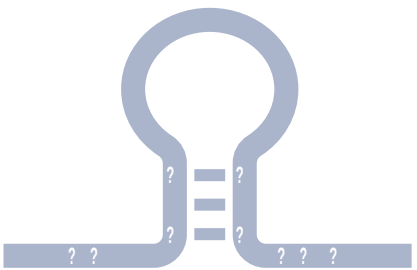
Data

terminator



+

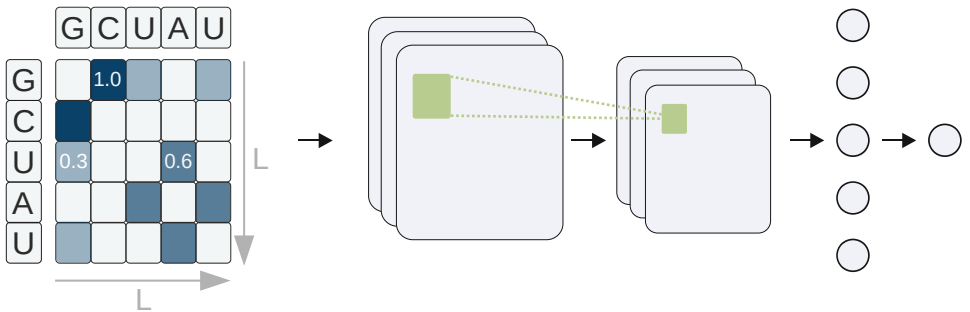
inverse-folded



Topology

matrix

2D-CNN

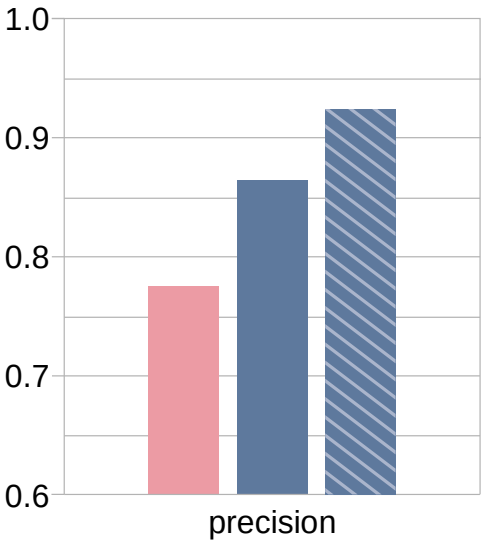


models

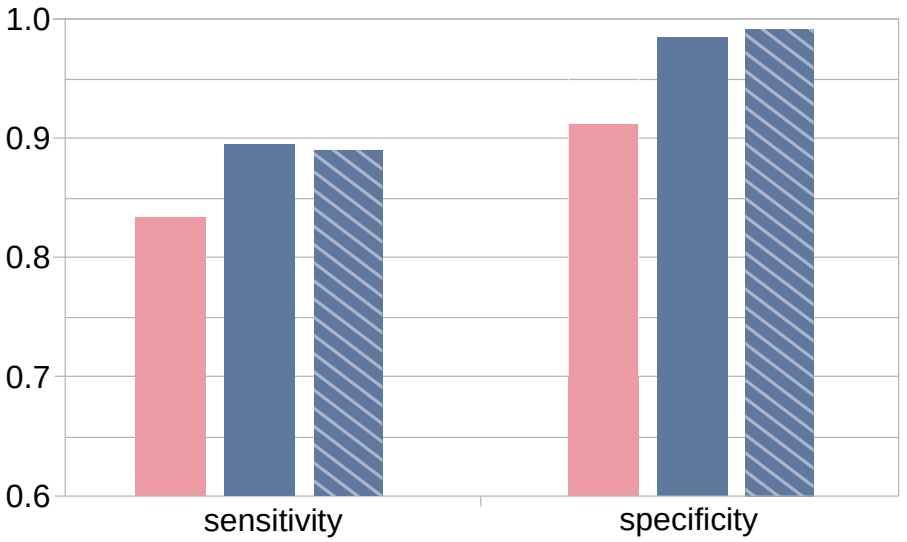
- matrix CNN
- matrix CNN pre-trained

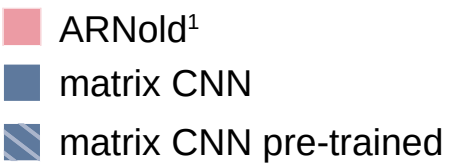
Performance

E. coli whole genome



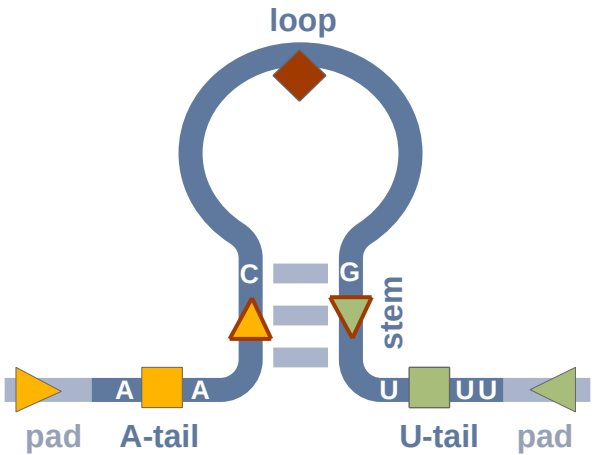
test dataset

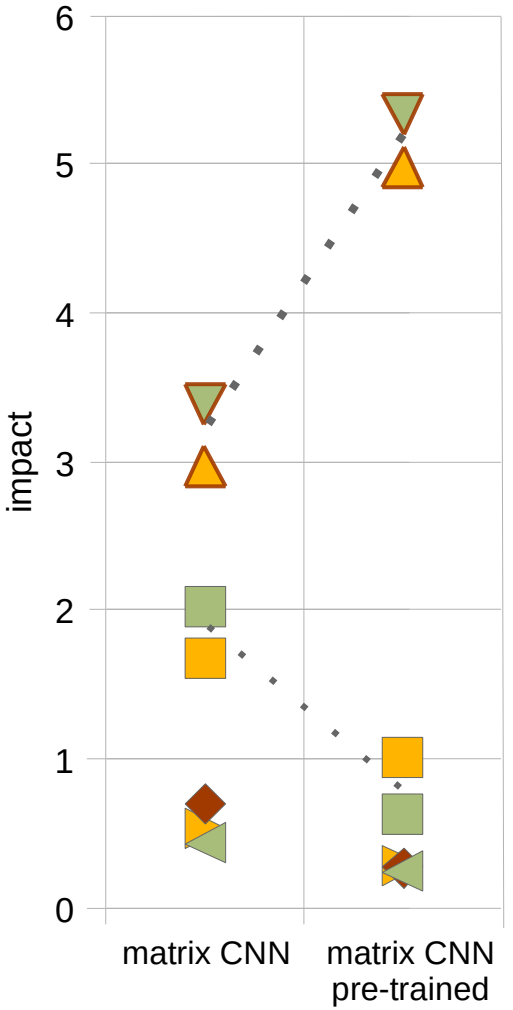




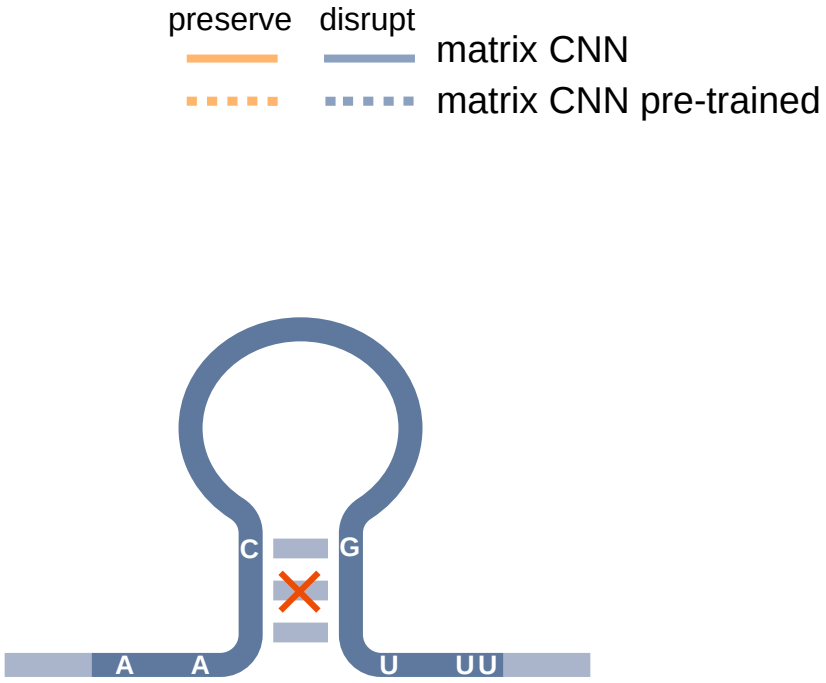
1: Naville *et al.*, RNA Biology 2011

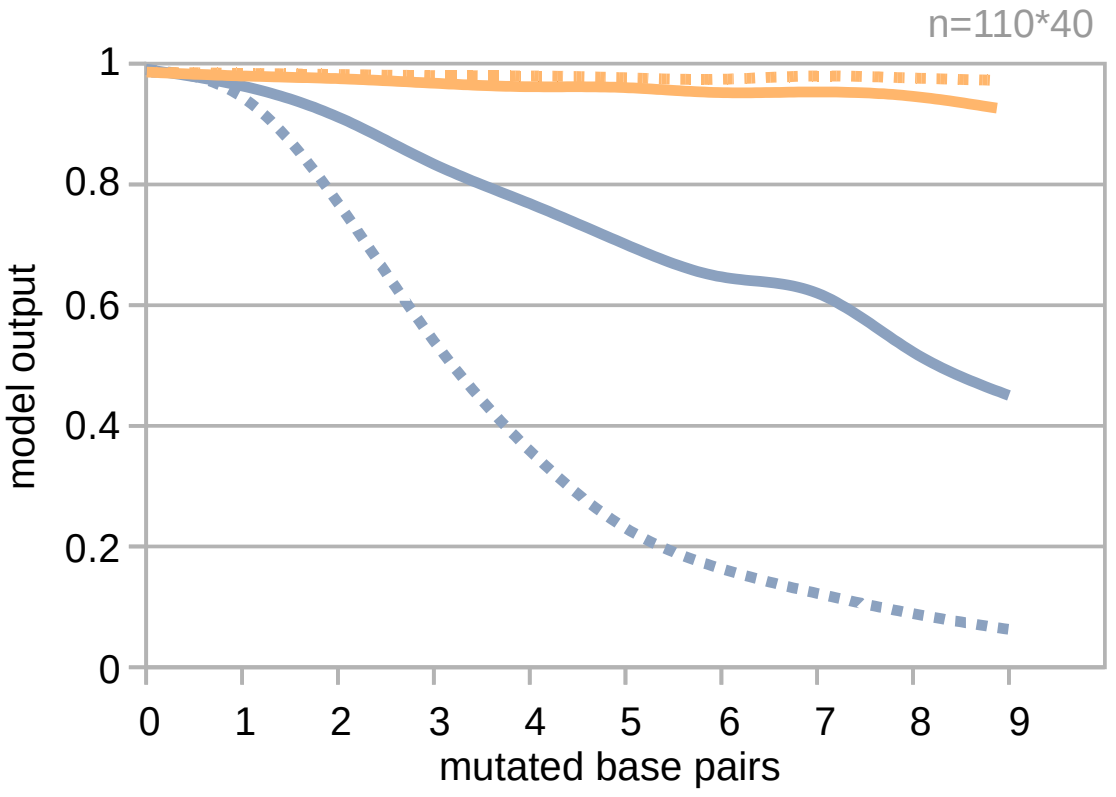
Impact: Sections



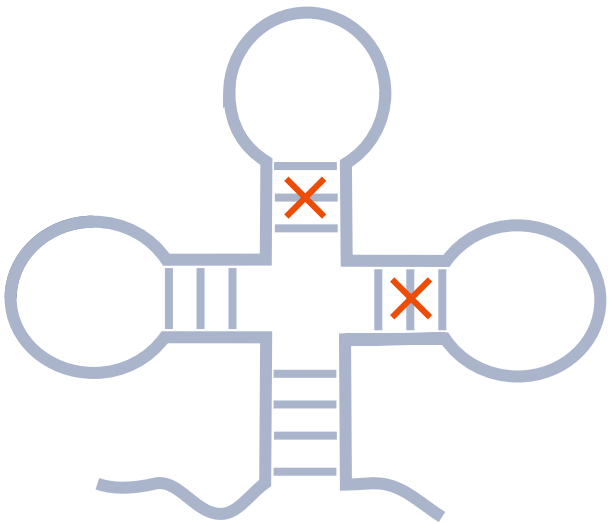


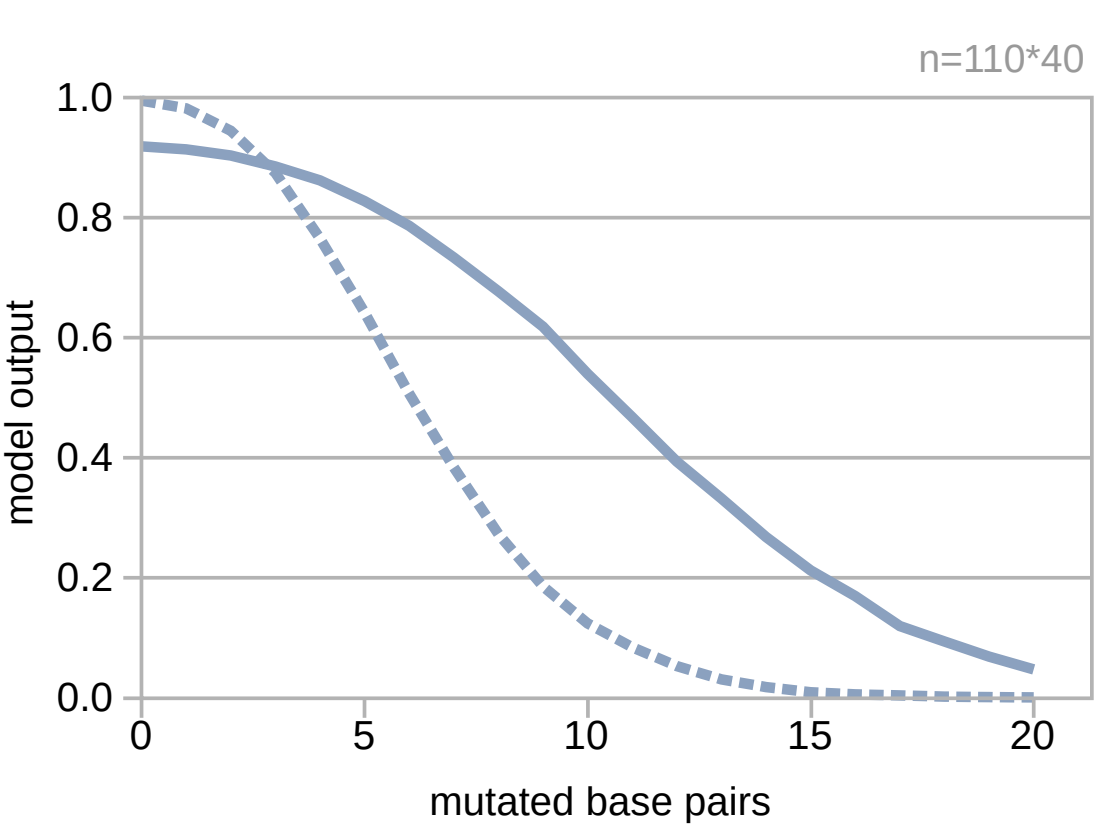
Impact: Base Pairs



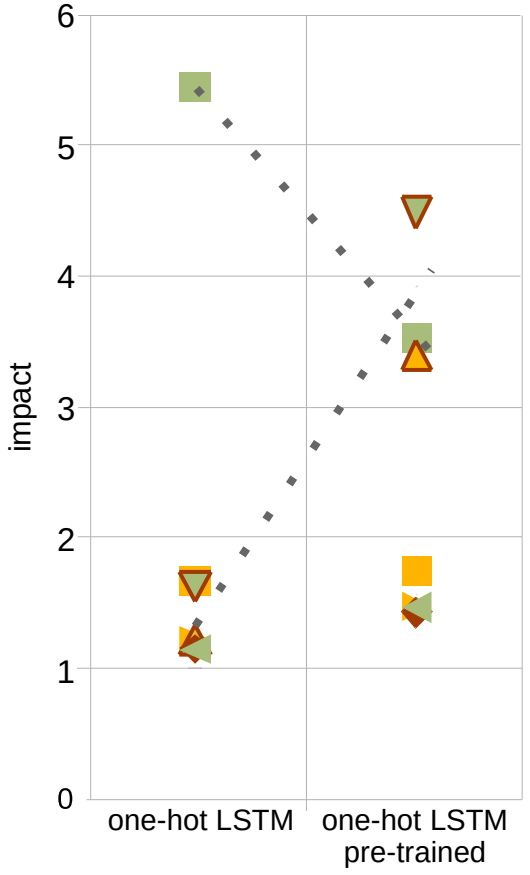


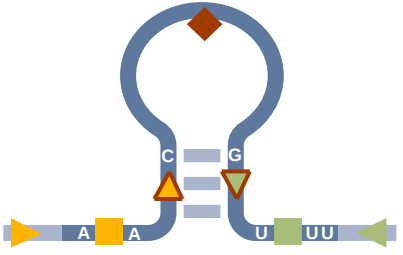
Application to other RNAs

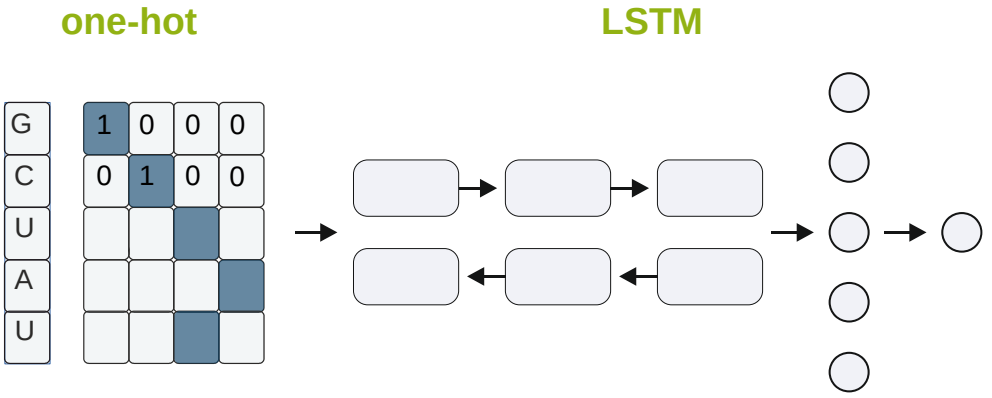


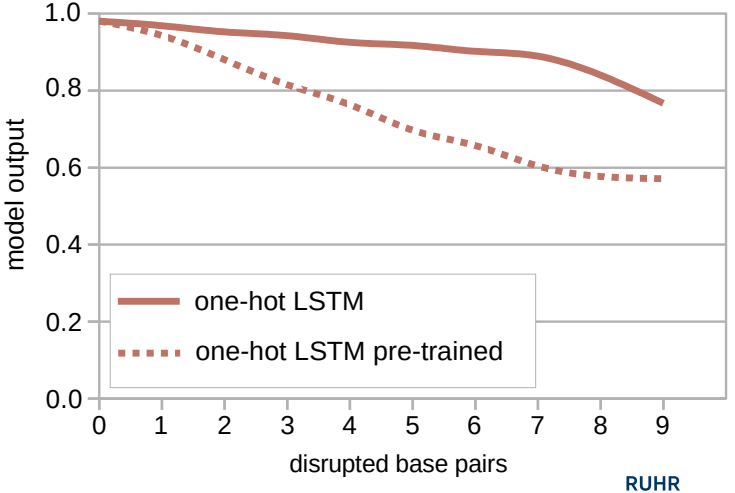


Application to other Topologies









Summary:

- Improves detection of Intrinsic Transcription Terminators
- Enhances learning of RNA structures
- Is applicable to other RNA families & different deep learning topologies
- Can be used to (partially) overcome limited data availability