# Prediction of Intrinsic Transcription Terminators using Deep Learning Methods

<u>Vivian B Brandenburg</u>, Franz Narberhaus, Axel Mosig; Ruhr-University Bochum, Germany
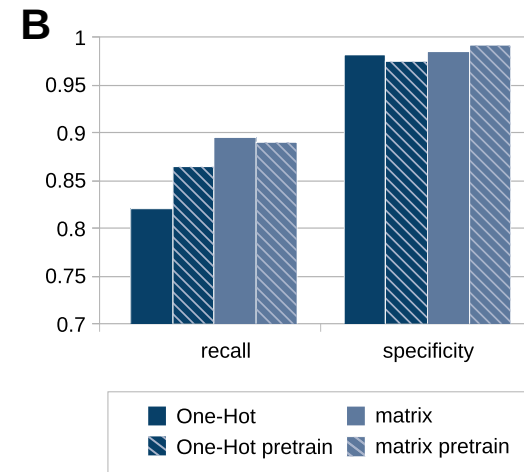
Intrinsic Terminators

Intrinsic Terminators form on the 3'-end of nascent RNAs, where they initiate the termination of the transcription elongation. They consist of a simple hairpin structure, followed by a U-stretch motif, and vary drastical in length.
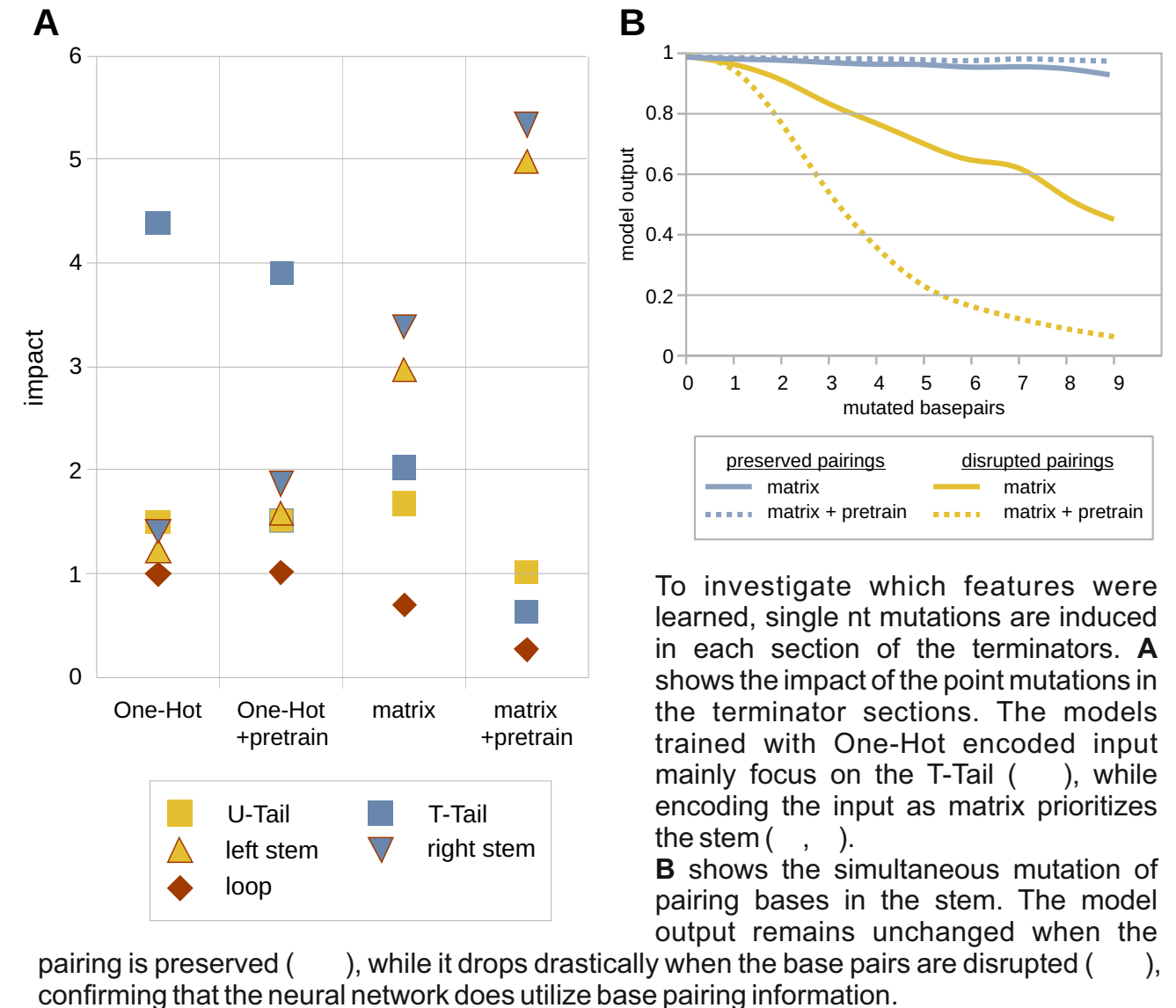

Method

A

structure-driven input

One-Hot encoding

terminator input
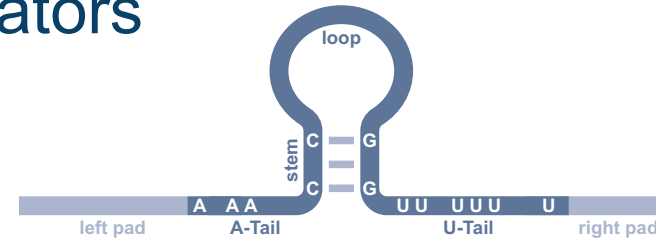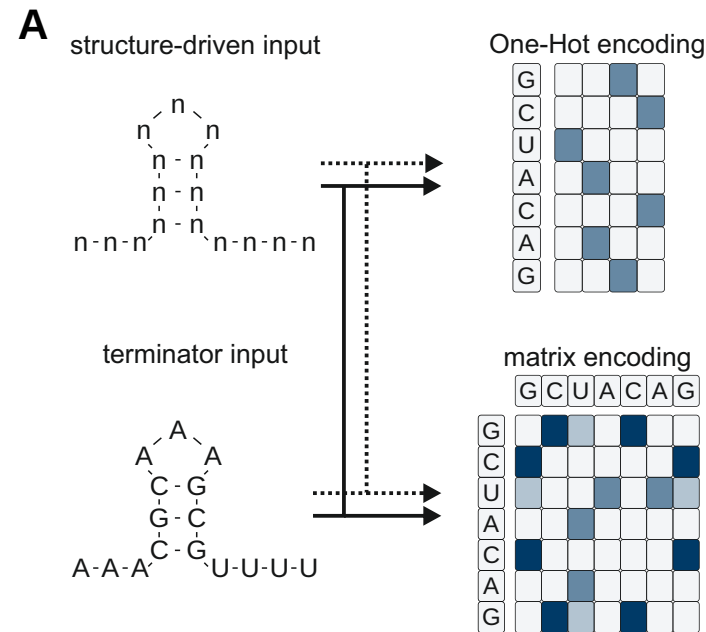
matrix encoding

Convolutional Neural Networks were trained with intrinsic terminators from *B. subtilis*[1] and *E. coli*[2]. **A**: We tested two encoding strategies: One-Hot encoding, where each nt is encoded as a 4-digit vector, and matrix encoding, in which possible base pairings are denoted.
The models were also trained with structure-driven input data, formed by inverse folding[3] of structures deriving from *B. subtilis* terminators.
B: All models learned to discriminate intrinsic terminators from other RNA sequences.

B

- One-Hot
- One-Hot pretrain
- matrix
- matrix pretrain

Learning of RNA structure & motifs

A

- U-Tail
- left stem
- loop
- T-Tail
- right stem

B

preserved pairings
- matrix
- matrix + pretrain

disrupted pairings
- matrix
- matrix + pretrain

To investigate which features were learned, single nt mutations are induced in each section of the terminators. **A** shows the impact of the point mutations in the terminator sections. The models trained with One-Hot encoded input mainly focus on the T-Tail (   ), while encoding the input as matrix prioritizes the stem (,   ).
B shows the simultaneous mutation of pairing bases in the stem. The model output remains unchanged when the pairing is preserved (   ), while it drops drastically when the base pairs are disrupted (   ), confirming that the neural network does utilize base pairing information.

Conclusion

Convolution-based models can not only learn to detect intrinsic terminators, but also learn the sequence motif and structure along the way. When trained with matrix encoded input, the model changes it's focus from the sequence motif towards the RNA structure. This effect is further enhanced when designed data are used for additional training.
In future, the implication of artificial data could also be beneficial to overcome the limited availability of known RNA structures for a wide range of RNA classes.

References

1. de Hoon *et al.*:Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS computational biology,* 2005
2. Chen *et al.*: Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods,* 2013
3. Lorenz *et al.*: ViennaRNAPackage 2.0. *Algorithms for Molecular Biology,* 2011